UNIVERSITY OF LISBON

Faculty of Medicine

Building-Up New Approach Tendencies in Individuals with High Versus Low Fear of Contamination

João Daniel Galvão Antunes

Supervisors: Dra. Lena Helene Ernst

Professor Dr. Tiago Vaz Maia

Dissertation elaborated to obtain the degree of Master in Neurosciences

2019

UNIVERSITY OF LISBON

Faculty of Medicine

Building-Up New Approach Tendencies in Individuals with High Versus Low Fear of Contamination

João Daniel Galvão Antunes

Supervisors: Dra. Lena Helene Ernst

Professor Dr. Tiago Vaz Maia

Dissertation elaborated to obtain the degree of Master in Neurosciences

2019

A impressão desta dissertação foi aprovada pelo Conselho Científico da Faculdade de Medicina de Lisboa em reunião de 19 de Novembro de 2019.

Acknowledgments

During the time dedicated to the development of the current work thesis, fear, uncertainty, lack of confidence and other issues arose. However, it is also important to mention those were not at all unexpected. Nonetheless, it was the individuals who surrounded and accompanied me during this time that allowed me to learn immensely a new scientific field, overcome the problems inherent of this choice and complete this journey in a successful manner. It is to them that I would like to dedicate the following kind words.

First, I would like to thank Prof. Tiago Maia for accepting me in his research group and whose enthusiasm about his research topics brought me into developing this thesis in his lab.

To my lab colleagues, I would like to thank their company in this journey: Catarina, for allowing me to share my passion for music, Vasco for being my snooker mentor. In particular, I truly feel in debt to Ana, whose kindness endured throughout this thesis in addition to helping in the design of the pilot study, and Angelo, whose lunch conversations were always fun, engaging and enlightening.

To my supervisor Lena, I absolutely need to thank the guidance, the advises and the confidence she placed on me during the thesis, as well as the calm and effective feedback that allowed me to incrementally improve the skills needed to complete this journey. To her, a big "danke sehr"!

To my school and university friends from Algarve, I would like to thank them for allowing me to every now and then share my journey with them and look back on the good times we had, even after our paths have diverged.

And of course, my family, specially my parents. This thesis was unquestionably completed with their support: I would like to thank my father for teaching me the discipline and responsibility necessary to stick with my goals no matter the struggles, and my mother for giving me the courage and love to take big leaps into the unknown. I owe this thesis to them.

Abstract

The daily life of patients with strong fear of contamination – as a sub-type of Obsessive-Compulsive Disorder (OCD) – is impaired by enhanced automatic avoidance tendencies. Standard treatment includes Exposure and Response Prevention Therapy, which is very effortful and results in a high rate of drop-outs.

The computerized Approach-Avoidance Task (AAT) might constitute an add-on therapy tool by building-up new connections between contamination-related stimuli (S) and approach reactions (R), that are less dependent on cognitive control. To avoid confounding effects by frequent comorbidities, two groups of healthy participants were pre-selected: 20 subjects with high (HG) and 21 subjects with low fear of contamination (LG) trained to approach contamination-related pictures with a joystick for 5 days in-a-row. Analyses were done by fitting a Power Law Curve and applying Mixed-Effects Models.

In line with the hypothesis of building-up new S-R connections, the LG mainly speeded-up the beginning of their reactions. In contrast, the HG decreased reaction times mostly after having initiated the response, but showed generally faster initiation times in the beginning, also for the control condition *avoid neutral*. This hints to heightened cognitive control in the HG throughout the training. In the ratings, specifically the trained pictures became less unpleasant from pre to post training. In a task version, where participants did not directly pay attention to the stimuli, an increase of approach tendencies from pre to post training was observed for the negative images in general, specifically for the LG. Groups did not change their reactions to untrained images of weak and strong content, nor did they differ in a practical test.

In the long-term, detailed information on optimal settings are indispensable to establish the AAT training as a powerful add-on therapy in OCD.

Key words:

Approach-Avoidance Task; Behavioral Training; Obsessive-Compulsive Disorder; Mixed-Effects Models; Experimental Psychology

Resumo

O dia-a-dia dos pacientes com forte medo de contaminação – um subtipo da Perturbação Obsessivo-Compulsiva (POC) – é afetado por elevadas tendências automáticas de evitamento. O tratamento usual inclui a Terapia de Exposição e Prevenção de Resposta, quer requer muito esforço e tem uma alta taxa de desistências.

A Tarefa de Aproximação-Evitamento (TAE) pode ser uma ferramenta de suplemento à terapia ao fortalecer as conexões entre estímulos (E) relacionados com contaminação e reações (R) de aproximação, que são menos dependentes de controlo cognitivo. Para prevenir o efeito de confundidores por comorbidades frequentes, dois grupos de estudantes saudáveis foram pré-selecionados: 20 sujeitos com elevados (GE) e 21 sujeitos com reduzidos (GR) traços de medo de contaminação treinaram a aproximar imagens com conteúdos de contaminação, usando um joystick durante cinco dias consecutivos. As análises foram feitas ao ajustar uma Power Law curve e aplicando Mixed-Effects Models.

Em linha com as hipóteses de fortalecimento de novas conexões E-R, o GR tornou-se mais rápido no início das reações. Em contraste, o GE diminui os seus tempos de reação após ter iniciado a resposta, mas demonstrou movimentos de iniciação mais rápidos no início, também para a condição controlo *evitamento do neutro*. Isto sugere um controlo cognitivo mais elevado pelo GE durante o treino. Nas classificações, as imagens treinadas tornaram-se menos desagradáveis de antes a depois do treino. Na versão de avaliação da tarefa, onde os participantes não prestaram atenção ao conteúdo dos estímulos, observando-se um aumento geral das tendências de aproximação de antes a depois do treino para as imagens negativas, especificamente pelo GE. Ambos os grupos não mudaram as suas reações para imagens não-treinadas de conteúdo fraco e forte, nem diferiam num teste prático.

A longo prazo, informações detalhadas das configurações ótimas são indispensáveis para usar os treinos da TAE como suplemento à terapia na POC.

Palavras-chave:
Tarefa de Aproximação e Evitamento; Treino Comportamental; Perturbação Obsessivo-Compulsiva; Mixed-Effects Models; Psicologia Experimental

Resumo Alargado

Os modelos baseados em dualidade de processos assumem que o comportamento humano é guiado por dois sistemas, nomeadamente o sistema refletivo e o impulsivo. O primeiro está maioritariamente envolvido durante o controlo consciente, enquanto o segundo está envolvido em comportamentos de hábitos e automáticos. No caso do sistema impulsivo, de um modo geral, estímulos benéficos provocam comportamentos de aproximação, enquanto estímulos prejudiciais provocam comportamentos de evitamento. A Tarefa de Aproximação-Evitamento (TAE) computorizada tem sido utilizada para o estudo desses comportamentos: os sujeitos fazem movimentos de aproximação e evitamento com um joystick o mais rápido possível, em resposta a imagens apresentados no ecrã de um computador. Tempos de reação (TR) mais rápidos indicam conexões preexistentes de estímulo-resposta (E-R) mais fortes, e, logo, tendências de reações automáticas mais fortes para um estímulo.

Os dois sistemas interagem durante a formação de hábitos, isto é, quando uma nova ação – inicialmente realizada para um determinado objetivo – progride para uma resposta que é automaticamente provocada por um determinado estímulo. Pensa-se que esta transição ocorre ao serem realizadas respostas motoras para um estímulo repetidamente, que por sua vez leva a um fortalecimento gradual das conexões de E-R. Tendo isto em conta, têm sido desenvolvidos protocolos de treino da TAE para fazer com que participantes adquiram novas tendências de aproximação ou evitamento, através do fortalecimento de novas conexões de E-R.

Em pacientes com doença mentais, os protocolos de treino da TAE também podem ser servir de contrapeso para as tendências de aproximação-evitamento excessivas e patológicas. Na Perturbação Obsessivo-Compulsiva (POC), existe um subtipo caracterizado por um elevado medo de contaminação e fortes tendências de evitamento. As opções de tratamento incluem a Terapia de Exposição e Prevenção de Resposta (TEPR), que consiste em confrontações repetidas com o estímulo evitado sem realizar comportamentos de limpeza compulsivos. Esta terapia é muito desconfortante, requerendo um esforço consciente significativo e resultando numa alta taxa de desistências. Neste sentido, os treinos de TAE podem constituir num possível suplemento a esta terapia, ao fortalecer novos hábitos de aproximação para o estímulo evitado, isto é, ao fortalecer tendências automáticas de aproximação que poderão ser capazes que contradizer as tendências automáticas de evitamento patologicamente elevadas, sem recurso a um esforço cognitivo.

A tese atual teve como objetivo chegar a um melhor entendimento dos processos cognitivos envolvidos em treinos semelhantes aos referidos acima. Para prevenir a influência de comorbidades frequentes na POC como confundidores, sendo a depressão um exemplo, dois grupos de estudantes saudáveis com traços elevados e reduzidos de medo de contaminação, respetivamente, foram recrutados. Estes foram submetidos a um protocolo de treino da TAE para a aproximação de imagens com

conteúdos relacionadas com contaminação (estímulos negativos). As hipóteses principais foram as seguintes: comparado ao grupo com traços de medo de contaminação reduzidos, os participantes com traços de medo de contaminação elevados (1) tiveram – ao longo do treino - TR mais lentos quando aproximavam os estímulos negativos e (2) mostraram uma menor diminuição – de antes a depois do treino - no quão desagradáveis as imagens eram classificadas e nas suas tendências de evitamento.

Os participantes foram selecionados através de um pedido online, que continha o questionário de auto-relato Inventário-Resumido de Obsessão-Compulsão para medir os seus traços de medo de contaminação. Baseado num limiar anteriormente definido, 20 participantes foram convidados para o grupo com elevado medo de contaminação (GE) e 21 para o grupo com reduzido medo de contaminação (GR). As suas saúdes mentais foram também avaliadas de um modo geral através do Inventário de Sintomas Curto. O design do protocolo de TAE consistiu em cinco sessões de treino durante cinco dias consecutivos, durante os quais os sujeitos aproximaram repetidamente estímulos negativos. Estes estímulos foram escolhidos de acordo com o estudo piloto da presente tese e consistiram em imagens de sanitas sujas com três níveis de conteúdo diferentes (fracas, médias e fortes). O afastamento de imagens de cozinhas-neutras foi utilizado no treino como condição controlo. Os TR ao longo de todas as sessões de treino foram ajustados a uma curva de potência (English: *power-law curve*) para obter informação dos declives destas curvas, isto é, do grau de alterações dos TR. Com isto, os TR de todo o movimento do joystick foram analisados bem como dos seus dois subcomponentes, isto é, o tempo que os participantes levaram a iniciar o movimento do joystick e o tempo que levaram até ao fim do movimento. Para avaliar mais detalhadamente as mudanças dos TR para os estímulos negativos, os participantes realizaram uma versão de avaliação da TAE antes e depois do período de treino. Aqui, além dos estímulos negativos e neutros, os participantes também aproximaram e evitaram estímulos neutros e positivos (cenários de ruas e praias, respetivamente). Em contraste com os treinos, onde os participantes prestaram atenção aos estímulos apresentados, nestas sessões de avaliação foi dada uma instrução para reagirem de acordo com a direção de uma seta. Isto fez com que esta versão medisse a influência do estímulo na reação, no que diz respeito a características semelhantes a um hábito. Adicionalmente, foram obtidos os dados relativos ao quão desagradáveis/agradáveis os participantes classificaram as imagens apresentadas. Após a última sessão de treino, estes realizaram um teste prático para avaliar as suas tendências de evitamento para estímulos relacionados com contaminação numa situação semelhante a um contexto do dia-a-dia: oi registado o tempo que levaram a sentar-se numa almofada que exibia uma imagem de uma sanita suja. A análise estatística dos TR e das classificações foi efetuada com Modelos de Efeitos Misturados (English: *Mixed-Effects Models*), dadas as vantagens na separação ótima entre variâncias aleatórias e sistemáticas em designs com medições repetidas.

Quando aproximando estímulos negativos ao longo do treino, em linha com as expectativas, o GR acelerou os TR de todo o movimento do joystick mais do que o GE, e mais do que na condição controlo. Estes efeitos foram impulsionados pelos tempos de iniciação. Por outras palavras, o GR diminuiu principalmente o início das suas reações de aproximação para os estímulos negativos, o que pode indicar que estes fortaleceram novas conexões de E-R como pretendido. Em contraste, foi encontrado o padrão oposto ao analisar o fim das reações: o GE demonstrou uma diminuição mais forte dos TR do que o GR, e mais forte do que para a condição *evitamento do neutro*, isto é, o GE diminuiu os TR para os estímulos negativos principalmente depois de ter iniciado a reação. Ainda de salientar, o GR demonstrou em geral tempos de iniciação mais rápidos no início, independentemente da condição. Isto parece indicar que o GE tinha níveis de controlo cognitivo mais elevados durante todo o treino, enquanto tentava acabar mais rápido a condição para aproximar os estímulos negativos.

Ao analisar os TR das aproximações *versus* evitamentos na versão de avaliação da TAE, isto é, quando os participantes não prestavam diretamente atenção aos estímulos, foi observado um aumento nas tendências de aproximação, de antes a depois do treino, para as imagens negativas em geral. Este efeito foi mais pronunciado no GR do que no GE. Ambos relataram que as imagens negativas usadas no treino tornaram-se menos desagradáveis de antes a depois do treino, do que as imagens não treinadas (ambas de conteúdo médio). Estes resultados estão em linha com a suposição geral de que é necessária uma consciência inicial para desenvolver uma nova resposta habitual: as classificações relatadas conscientemente mudaram depois do treino especificamente para as imagens treinadas, enquanto que o período de treino pode não ter sido suficientemente longo para alterar as tendências de reação de um modo já tão específico. Análises posteriores não revelaram evidências para efeitos do treino nas imagens negativas de conteúdo fraco nem para as de conteúdo forte, ambas que não tinham sido utilizadas durante o treino. Aqui, é importante referir que ambos os grupos classificaram as imagens fortes como sendo mais desagradáveis do que as fracas, no qual o GE classificou todas as imagens como sendo mais desagradáveis do que o GR, em geral. Por último, no teste prático, não houve diferenças entre grupos relativas ao tempo até sentar na almofada com a imagem negativa.

Tomados em conjunto, estes resultados revelam dados importantes no modo de como os diferences processos cognitivos envolvidos na formação de hábitos são influenciados pelo treino na TAE. Análises futuras irão implementar modelos de aprendizagem da psiquiatria computacional para melhor separar as influências do controlo cognitivos e aprendizagem habitual, ao permitir quantificar os parâmetros que não são diretamente observáveis no comportamento. A longo prazo, estes resultados permitirão revelar o modo como os protocolos de treino de TAE podem ser concebidos como suplementos à terapia para compensar dificuldades nas intervenções de tratamento padrão para a POC.

List of Contents

List of Figures

1.Introduction

Content:

- ✓ ***Fundaments of Human Behaviour***
- ✓ ***Obsessions and Compulsions***
- ✓ ***Current Thesis***
- ✓ ***Hypotheses***

1.1 Fundaments of Human Behaviour

1.1.1 Reflective-Impulsive Model

Research and analysis of human behaviour has led to the proposition that it is guided by two distinct but interchangeable systems, as suggested by some dual-process models, that try to explain differences between impulsive (automatic) and reflective (conscious) processes[1]. The Reflective-Impulsive Model is one of these models, according to which behaviour is the result of an interplay between two systems, namely the Reflective and the Impulsive systems[2].

As initially proposed by Strack & Deutsch[2], the two systems operate to elicit behaviour, whereby in the Reflective System "*behaviour is the result of a decision that is based on the assessment of a future state in terms of its value*". In contrast, the Impulsive System "*elicits behaviour through the mere spread of activation*" of associated contents "*by motivational orientations*". Based on these definitions, Strack & Deutsch characterized the former system as ensuring flexibility and operating slowly, whereby its processes depend on intention. As for the latter system, information is processed automatically throughout associated contents (i.e., response and contextual cues), whereby the links between them are slowly formed and strengthened over many learning trials. Along with this idea, other studies also supported these characterizations: decision making and self-regulation, both of which being higher cognitive functions that are based on the assessment of a future state have a limited capacity[3], while the automatic processing of context stimuli can results in responses with no prior intentions[4,5].

The interaction between the two systems is dynamic, wherein both can influence each other independently[2]. For instance, the Reflective System can be influenced by impulsive factors, such as frequent activation and motivational orientation, which then can alter the accessibility of conceptual contents during reasoning. On the other hand, the Impulsive System can be influenced by self-regulation strategies which delay gratification from tempting stimuli[2]. Lastly, both systems can act together, whereby the reflective-derived processes can progressively turn to implicit, automatic behaviour, upon being repeatedly linked to stimuli[6,7].

1.1.2 Stimulus-Response Connections

One of the interactions between the Reflective and Impulsive system that is of most interest for the current thesis is the way in which reflective-derived behaviour transitions to implicit responses during acquisition of new behaviours, a process that encompasses goal-directed behaviour and habit learning, respectively. The former is cognitively more effortful and sensitive to the value of prospective goals, being involved

during the acquisition of new behaviours and guided by the expected value of the final outcome.

On the contrary, a habit is an automated response which is strongly and rigidly tied to the context of the performance, such as environment and preceding actions in a sequence, as a result of progressively acquired associations between the response and the stimuli [8–10]. According to this definition, and considering the context of the thesis, one can view habits as the outcome of *implicit goals* and *direct cueing*. The first term refers to the initial habit formation that occurs when repeatedly chasing a goal via a specific behaviour in a specific context, until the performance itself becomes goal-independent[10]. Linked to this, the second term refers to the consequent cognitive neural changes that result from the repeated co-activation of stimuli/context and associated responses[11], and whose brain correlates are mentioned in section *1.1.4 Brain Correlates* below.

One additional important dimension of goals and habits is how they differently interact depending on the situation at hand[11]. Of importance to this thesis is their interaction during behavioural acquisition of new habits, which occurs through repetitive motor responses initiated in response to a specific stimulus, leading to an incremental strengthening of associations between the stimulus and the executed response, i.e., stimulus-response (S-R) connections. As a result, there is a gradual shift from a goal-directed response to a habit, wherein the conscious behaviour become increasingly implicit and automatic upon contact with the specific stimulus [8,9,11]. In this perspective, habitual responses operate in the service of goals, whereby the second directs control of response prior to formation of the first, therefore defining the contextual cues under which a response is repeated to become a habit[11].

1.1.3 Approach and Avoidance Behaviours

As mentioned above, the Reflective-Impulsive duality seem to guide actions and the two main processes that are involved in the acquisition of new behaviours. Yet, a more fundamental distinction of behaviour types can be made, namely approach and avoidance responses and their relation to valence of a stimulus, which also influence habit formation. Evidence for this basic distinction of human behaviour is supported by evolutionary, behavioural psychology and neurophysiological findings[12]:

A significant number of studies have shown that humans have the capability to evaluate most stimuli they encounter, such as facial expressions[13], words[14] and others[15], in a good-bad dimension[16] even without intention[17], triggering certain behaviours unconsciously[14,18]. With this knowledge as a background, theories that link stimuli valence to behaviour have suggested that stimuli evaluated as displaying positive characteristics, i.e., associated with a beneficial outcome, seem to activate specific neural circuits that trigger approach-related behavioural tendencies, while stimuli associated with negative characteristic and a harmful outcome tend to activate neural circuits that trigger avoid-related behavioural tendencies[2,19,20]. This valence-behaviour

compatibility led some researchers to suggest a direct response to stimuli connection unmediated by cognitive processes[21,22]. In fact, these evidences were later used to propose the idea of a bidirectional link between stimulus valence and behaviour, further developed by the first proponents of the RIM[2]. According to it, not only the valence of the information processed in a stimulus facilitates a compatible behaviour, but the behaviour executed could itself exert an influence over the stimulus evaluative processes.

Additionally, evidence from evolutionary biology suggests that approach and avoidance-like behaviours are observed in a variety of organism besides humans[23] and appear to be pre-programmed to particular classes of stimuli[24], possibly due to their contribution to the organism's environmental adaptation[25,26]. For example, positively-valenced stimuli, i.e., whose outcome became associated with beneficial effects due to previous experiences, tend to facilitate behaviours towards that stimuli. On the contrary, negatively-valenced stimuli, i.e., whose outcome is associated with harmful consequences for the organism, tend to facilitate behaviours away from the stimulus[18,27].

Lastly, neurophysiological evidence has also supported the approach-avoidance distinction, by showing identifiable neuroanatomical structures and neurotransmitter activity underlying approach and avoidance-related behaviour. These are further explained in section *1.4*.

Given all these evidences for approach and avoidance behaviour, the explanation as to how these responses are dependent on stimuli valence and contextual factors is provided in the section *1.1.5 Approach-Avoidance Task*.

1.1.4 Neuronal Correlates

1.1.4.1 Goal-directed Behaviours and Habits

The need to explain complex behaviours in behavioural research led to the proposal of Thorndike's Law of Effect, according to which the repeated responses that provide a sense of satisfaction in a given situation would be more firmly connected with the context, so that when similar conditions would occur those responses would be more likely to happen[28]. In practical terms, this means that beneficial outcomes strengthen the S-R connections, while harmful outcome have an opposite effect. Afterwards, with research showing that animals could control their actions according to the anticipation of the outcome[12,13,29]. In the reinforcer devaluation paradigm, for instance, an animal with extended training for pressing a bar for food would continue to perform the habitual response even if the food reinforcer was substituted with a toxin, revealing a reinforcer devaluation insensitivity in the habitual response[11]. In dual tasks that involve motor learning, evidence shows that extended practice of an audio-vocal two-choice RT task (identifying an audio tone as being "high" or "low", depending on its frequency), while simultaneously performing a discrete motor task (replicate the trajectory shown on-screen, with a vertical lever), leads to improved performance in

both tasks[30]. As such, this suggests that as practice progresses less attention is required for the tasks, which is one of the characteristic steps during habit formation.

Further characterization of habitual compared to goal-directed behaviour development was possible due to electrophysiological recordings and lesion studies. The first showed that activity in striatum cells could predict performance improvement throughout training[31–34]. To pinpoint more precisely how the striatum processed habit acquisition, lesions studies showed that the dorsomedial striatum in mice (equivalent to the caudate in humans) was involved in initial flexible choice behaviour to a newly-set goal, as well as acquisition and expression of habitual behaviour, whereas the dorsolateral striatum (equivalent to the putamen in humans) was associated in maintaining the habitual response upon extended training. Indeed, electrophysiological recordings also supported this evidence, showing a transfer of neural activity from dorsomedial to dorsolateral striatum over the course of training[35,36] (for an extended review see [29,37,38]). At the synaptic level, electrophysiological recordings also revealed that the neurotransmitter dopamine played a major role in the long-term synaptic plasticity that underlies leaning processes (Long-Term Potentiation) in the striatum, since its activity is sensitive to reward: the dopamine release in the synapses promotes the strengthening of neuronal connections upon encountering an unexpected reward. Through this biological mechanism, the organism is more likely to repeat the behaviour that leads to the reward, endorsing S-R reinforcement in the long-term[39,40].

Advances in neuroimaging procedures, where the neural correlates of task performance (such as outcome devaluation and learning tasks) are monitored during training, allowed to further support the shift away from goal-directed control to habitual responses over the course of a training. Most of these studies usually report a reduced activation of the prefrontal cortex (whose functions include cognitive processes like organization and planning[41]) and increased activity of subcortical structures[11]. In particular, studies that specifically look at striatal activation show, despite of some task-related differences, activation of putamen regions during habitual responses[37,42,43], which is consistent with the electrophysiological recording and lesions studies above. Lastly, with regards to dopaminergic activity, neuroimaging has also shown that the ventral striatum (which encompasses the dopaminergic system) is associated with rewards obtention and learning[44]. In fact, given the reward prediction property (i.e., the difference between the predicted and actual received reward) being encoded by dopamine activity for unexpected rewards, one can also evaluate habit formation by monitoring when the prediction error approaches zero[45]. Indeed, this reinforcement learning measure has been used to predict, although only to some extent, neuronal activity and behaviour[44,45].

Despite identifying the behavioural and neural correlates of goal-directed and habitual behaviours, the definition of the features that characterize habit learning, such as slow, unconscious and automatic, are hard to accurately define and standardize for different tasks since they depend on the nature of the task itself. For instance,

determining if striatal dependent learning happens slowly, i.e., through incremental changes that would reflect the S-R connection strengthening, varies in how many trials it takes to reach maximal performance. In addition, with regards to consciousness, assessing awareness during task performance can affect the strategic approach to the task itself. Lastly, habit inflexibility is usually defined when comparing the response flexibility enabled by executive functions (for an extended review see [45]). Thus, when trying to classify a response as goal-directed or habitual, one must not only consider the theoretical definitions but also consider the context in which the behaviour is performed given the nature of the task.

1.1.4.2 Approach-Avoidance Behaviours

Concerning the neural structures underlying approach and avoidance, neuroanatomical research in animal models and human neuroimaging studies have evidenced a cortico-striatal circuit, encompassing structures such as the pre-frontal cortex, the anterior cingulate cortex, the striatum, the amygdala and the insula (for an extended review see [46,47]. Based on these studies, a neurobiological triadic model was developed as an attempt to explain approach-avoidance behaviour in contexts where the pre-frontal cortex has a suboptimal functioning, such as in adolescent behaviour and psychopathological conditions. To this end, the model addresses three neural systems[48]: the approach system, that underlies actions in response to reward cues; avoidance system, responsible for withdrawal in response to aversive cues; and the regulatory control system, which provides a modulatory control of the approach and avoidance behavioural systems. In turn, these systems would be guided by three distinct set of brain circuits: the ventral striatum as the main region for processing approach, reward-driven behaviour; the amygdala as mediator of avoidance behaviour; and the pre-frontal cortex as the hub for cognitive control[48].

Given the oversimplification of behaviour attribution and information processing to specific brain circuits, an improved version of the model was developed[49]. This model extension was conceived so that each of the three structures itself was a triadic formation in order to take into account the different roles, although to a minor degree, each structure evidenced other than its main function.

1.1.5 Approach - Avoidance Task

The first studies that investigated the link between stimuli appraisal and responses demonstrated evidence for an automatic compatibility between the two, that is, without conscious awareness[14,17,21]. This compatibility was best shown in what has been termed the "*specific muscle activation hypothesis*"[50]. According to this hypothesis, a hardwired link between stimulus evaluation and motor activity was responsible for driving automatic arm flexion (pulling) towards positive stimuli (something considered desirable) and arm extension (pushing) in relation to negative stimuli (something considered aversive)[21,22]. Despite this initial idea, later evidence showed that arm flexion

and extension can depend on situational contexts, whereby the same arm movements can result in different outcomes (reviewed in [18]).

Markman and Brendl[51] provided one of the first evidences that contextual factors play a role on approach and avoidance behaviours, using a design that varied the location of the "*self*" in relation to a positive or negative attribute. More precisely, the task had each participant's name displayed on a virtual corridor shown on a computer screen. Positive and negative words were presented either more distant or nearer than the names along the corridor, whereby participants had to push away or pull closer a vertical lever in order to move the words closer to their names. Results showed that participants were faster at moving positive words towards their names and faster at moving negative words away from their names, regardless of arm movements. This suggested that the compatibility effect was dependent on participants' perception of their selves, rather than simply of their actual physical location[51]. In another study[52], participants had to react to positive and negative words using a joystick, in which the points of reference was either the actual self or the stimuli shown in the computer. Results showed that in an actual self-point of reference, reactions to positive words were performed faster when pulling the joystick towards the body, while reactions to negative words were faster when performing pushing the joystick away from the body. In contrast, in a stimuli-point of reference, reactions to positive words were faster when pushing away the joystick from the body, while reaction to negative words were faster when pulling towards the body. In line with these results, other studies have also provided additional evidence that behaviours depend on contextual factors[53,54].

To study the hypothesis on appraisal-arm movement compatibility in more detail, most studies have used the Approach-Avoidance Task (AAT), in which participants' approach-avoidance behaviour towards a set of stimuli is evaluated. In this task, to ensure that implicit processes are measured, the experimental setups are designed in a way that subjects are instructed to react as fast as possible to stimuli presented on a computer screen, by using a joystick as a means to approach (pull motion) or avoid (push motion) the stimuli on the screen. More importantly, participants are instructed to pay attention not to the content of images but rather to a "*task-irrelevant*" feature such as frame colour or direction of an arrow, which is shown together with each image. With this framework, the reaction-times (RTs) for the joystick responses given to each stimulus are registered. Different iterations of this task have been developed, namely the Manikin Task, the Joystick Task, and the Feedback-Joystick Task. In the first, using keyboard button presses, participants control the movement of a manikin on a computer screen relative to the position of the stimulus. More precisely, the manikin randomly appears around the stimulus, for example, right or left of the stimulus, and subjects then have to move the manikin away or closer to the stimulus, reflecting an approach-avoidance behaviour[55]. In the Joystick Task, a vertical joystick is used to reflect the approach-avoidance behaviour through the motions of pulling or pushing, respectively, relative to an image shown on the computer screen[21].The

Feedback-Joystick Task can be considered a modified version of the Joystick Task, whereby its zooming feature enables the gradual increase or decrease of the images while the joystick is being pulled or pushed, further enhancing the sense of approaching or avoiding, respectively, the stimuli in a real-life context.[56](for a detailed review on each AAT iteration see [18]).

In addition, two types of instructions can be provided to the participants, in an explicit or implicit manner. With explicit instructions subjects are instructed to pay attention and to react (approach or avoid) based on the content of the stimulus (and therefore to its negative or positive valence). In contrast, with implicit instructions subjects do not pay attention to the content of the stimulus but instead are instructed to react to a task-irrelevant feature.

Given these features, several studies have used the AAT as a means to not only assess automatic approach-avoidance behaviours, but also to modify them. These will be mentioned in the following sections.

1.1.5.1 Assessment of Approach-Avoidance Reactions

With the aim of replicating the findings of Solarz[14], Chen and Bargh[21] asked students to classify words on a computer screen in a good or bad dimension, as quickly as possible. Based on their condition assignment, students would have to push or pull a lever when they judged the word to be good or bad (congruent condition), respectively, or to push or pull a lever when they judged a word to be bad or good (incongruent condition), respectively. In a second separate experiment, students were instructed to just react as quickly as possible by approaching and avoiding an equal number of times to the presence of words. Both experiments showed that the approach and avoid reactions were relatively faster for the positively and negatively-valenced words, respectively. Most importantly, the second experiment showed that this link did not depend of conscious awareness. Using a similar design as Chen an Bargh, another study also showed that participants reacted faster to approach, compare to avoid, positive words and faster to avoid, compared to approach, negative words, using a joystick[52].

In a similar fashion, Alexopoulos and colleagues[57] tested the link between automatic evaluation of words and behavioural tendencies, now in a Joystick-related AAT version. In this study, the authors asked students to quickly react to sadness and happiness-related words though a modified keyboard with only 3 keys, that is, a central key as a starting point together with one upper and lower keys, the latter two being more distant and required arm extension and flexion, respectively, analogous to a push and pull arm movements. Subjects were instructed to press the up or down button if a word appeared on the upper of down portion of the screen, respectively. In line with previous aforementioned studies, subjects were faster to approach and avoid happy and sadness-related words, respectively.

1.1.5.2 Modification of Approach-Avoidance Reactions

The evidence that the AAT is sensitive to automatic approach-avoidance led to the development of AAT computerized protocols, which require subjects to repeatedly react to stimuli in an opposite way of the automatic tendencies they have for the stimuli. Generally, such protocols usually include: (1) A computerized task in which participants are instructed to push or pull a vertical joystick as quickly as possible in response to images shown on a computer screen; (2) One or more AAT training sessions to repeatedly pay attention to a specific feature of the stimuli shown; (3) baseline and post-training assessment of subjects' implicit behaviour when reacting to the stimuli, which are obtained by analysing the RTs when avoiding *versus* approaching. More precisely, faster RTs to pulling motions compared to pushing motions indicate an approach bias, while faster RTs to pushing compared to pulling indicates an avoidance bias. By comparing the biases when reacting to the stimuli before *versus* after the training period, it is possible to analyse training effects; (4) Joystick movements performed as fast as possible in response to a task-irrelevant feature in the pre and post-training assessment sessions, in order to capture reaction biases (RBs) unmediated by cognitive control; (5) Explicit ratings to the stimuli, in order to obtain a more conscious evaluation to the stimuli.

In the first studies that used an AAT training protocol to modify automatic reactions in healthy participants, Huijding and colleagues investigated the effects of manipulating fear-related tendencies in children[58,59]. In both of their studies, using a Feedback Joystick and a Manikin AAT version for the first and second study, respectively, children were assigned to either repeatedly move towards animal 1 and move away from animal 2, or the reverse**.** Assessment sessions consisted in discriminating animal 1 from animal 2 by pronouncing as fast as possible their respective name *and then* moving the joystick/manikin. Results of both studies showed that when comparing reactions between assessments, children made fewer approach movements towards the animal they had been assigned to avoid. In addition, regarding global attitudes (positive or negative) towards each animal measured before and after training, children reported an increase in self-reported liking of the animal they approached and a disliking of the animal they avoided. Additionally, in one of these studies[59], the authors added a post-training session where children decided for which animal he/she would like to know the answer to positive, negative or neutral questions. Interestingly, this showed that children asked more negative information about the avoided animal and more positive information about the approached animal.

Concerning emotional expressions, a sample of college students were instructed to push away or pull closer, images of neutral, angry and smiling faces shown in separate blocks, through a Joystick AAT [60]. Here the task-irrelevant feature was colour, whereby the instruction either was to push away images tinted with brown and pull closer images tinted with blue, or the reverse. To assess implicit face evaluations, an affective priming task was used. In this task, using keyboard keys, subjects had to categorize positive or

negative words as pleasant or unpleasant, whereby immediately before each word was presented a priming picture (angry, smiling or neutral) was shown for 300ms. Consequently, there were two picture prime-word combinations: *pull* a *positive* word preceded by a (neutral, angry or smiling) face previously trained to *approach* (congruent combination), or *pull* a *negative* word preceded by a (neutral, angry or smiling) face previously trained to *approach* (incongruent combination). The main finding was that in the affective priming task, subjects where faster when performing congruent trials than incongruent. However, this effect was observed only when the priming pictures were neutral faces and not in angry or smiling faces, for which the authors argued that it was due to non-ambivalence in angry and smiling expressions or a weak training effect which did not affect the strong valence of the angry and smiling faces. In a similar design but using neutral faces as the task-irrelevant feature, subjects were instructed to push away or pull closer, depending on the background colour, images with neutral faces[61]. Results showed no effect of approach-avoidance training on implicit face reactions, although the small effect sizes in the expected direction hinted the authors that more subjects would be needed to find a significant training[61].

To investigate the extent at which AAT training can modify individuals' automatic approach-avoidance tendencies on eating behaviour, three studies have been performed so far[62–64]. In the study develop by Becker *et al.* [62], using three separate designs, female students were randomly assigned to either official training or sham training, where they had to mostly (90% of the trials) avoid images of unhealthy food or avoid and approach these images in an equal frequency, respectively. More precisely, using two (right *versus* left) buttons on top of a stick, the first of the three designs included a training protocol that required participants to either avoid unhealthy food (experimental training) or to approach and avoid equal number of unhealthy and healthy food (sham training). In addition, participants performed a pre and post-training assessment sessions, where they approached and avoided an equal number of unhealthy and healthy food images. The other two designs had some different methodological aspects, including response using keyboard keys, increased sample size, an affective priming tasks similar to a previous study[61], explicit ratings of images and a behavioural test, which required subjects to eat and taste chocolates. In all three designs, no conclusive evidence was found regarding changes in implicit or explicit food preferences and eating behaviour[62]. With the same aim and similar training protocol applied to overweight children and adolescents as in [58,59] , the study by Warschburger *et al.*[64] used a Joystick AAT version to instruct a pulling motion in response to images of vegetables and a pushing motion in response to snacks. Results showed a significant change in automatic approach-avoidance tendencies, indicating that training was more effective in increasing avoidance of snacks compared to increasing approach of vegetables. Lastly, the study by Schakel and colleagues[63] also investigated the effects of food-related outcomes but through a gamified training. Here participants were randomly assigned to one of four conditions: Training with food-related games, all of

which were approach-avoidance related, including a Feedback Joystick AAT version; Training with food-unrelated games; Verbal suggestions-only, where additional information about the games and performance expectations would be provided, but no games would be played; And a combination of food-related games with verbal suggestions. Results showed that for participants that had been assigned to train a gamified protocol with verbal suggestions, subjects reacted faster for associations between healthy food and positively-valenced words and chose healthy compared to unhealthy food pictures in a food choice task after training, compared to the other training conditions.

To investigate if approach-avoidance training could modify automatic approach tendencies towards smoking, one study[65] instructed smoking individuals to push away smoking-related images and to pull closer neutral images, with a computer mouse. Results showed that training led to a reduction in smoking (comparison of participant's information about cigar and alcohol consumption from pre to post-training), jointly with a decrease in cigar dependence and compulsiveness, compared to a waitlist control group who had not trained. For a more detail review on AAT training studies to reduce smoking behaviour see [66].

Thus, evidence so far suggests that AAT training protocols, although not showing clear training effects on all fronts, have the possibility to modify up to a certain degree the behavioural responses for the trained stimuli.

1.1.6 Approach-Avoidance Measurements in Clinical Psychology and Psychiatry

Different types of motivational approach and avoidance imbalances have been associated with the development of distinct psychopathological disorders[67]. As such, further understanding of these disorders in a research setting usually involves assessing the underlying motivational approach-avoidance systems that may be driving the pathological behaviour. Prior to the introduction of the AAT [56,68], this was usually done by performing measurements on a trait behavioural level, through the use of self-report questionnaires which are assessed by clinicians. However, the introduction of the AAT as a more implicit way to evaluate the approach-avoidance systems motivated a variety of studies to use the AAT in psychiatric diseases. In particular, the hypothesis was that by looking at response latencies to disorder-relevant stimuli one could further evaluate and investigate the uncontrollable and automatic behaviour involved in a range of psychopathologies. Indeed, several studies have used the AAT to measure automatic tendencies in addiction and anxiety-related disorders, using the same features mentioned in the subsection 1.5.

In the context of Social Anxiety Disorder (SAD), Heuer *et al.*[68] showed images of emotional facial expressions (angry, smiling, neutral) and puzzles to highly-anxious (HA) and low anxious young adults (LA). Subjects were asked to react to the structure of the

images (faces or puzzles), rather than to their content and afterwards rated the images. Results showed that the HA group displayed stronger avoidance tendencies than the LA group for both smiling and angry faces. Results for the angry faces were consistent with the idea that angry faces convey information regarding potential threat, activating avoidance mechanisms[69]. As for the smiling faces the authors argued that it could be due to the HA's concern of being disliked and rejected, since to a socially anxious person even a friendly smile could be associated with threat. Despite the RTs differences between groups, there were no explicit rating differences between them[68].

In a similar procedure, one study[70] that used images of crowds displaying different emotional expressions (angry, happy or neutral) at the same time, social-anxious participants (SA) and non-social-anxious controls (NSA) were instructed to react to the colour of the image filter, rather than to the content of the images and to rate the images afterwards. Results showed that SA displayed a marginally stronger avoidance tendency when the number of angry faces in a neutral crowd increased, and generally avoided crowds of happy-neutral faces, compared to NSA controls who did not show any action preference for any crowd type. The authors suggested that a higher sensitivity to socio-evaluative threat in SA, compared to controls, was driving the general tendency of SA to push away social crowds, with faster avoidance tendencies as the threatening faces in the crowds increased. In addition, image ratings did not differ between groups.

Further extending these results, another study[71] investigated the influence of gaze direction on approach – avoidance responses, reporting that HA students were faster in avoiding angry faces, compared to LA. Moreover, this tendency was present only when the face's gaze was directed towards the subjects. In contrast to the initial hypothesis the HA group also displayed avoidance tendencies for happy faces for both gaze directions. These results were in line with previous research, showing that direct gaze is upsetting for anxious individuals[72,73], making these individuals generally avoidant of human interaction particularly in the context of angry faces.

So far, few studies have used the AAT paradigm to investigate automatic approach-avoidance tendencies in anxiety-related clinical samples. One of these[74] investigated approach-avoidance reactions to ambiguous and negative emotional faces in individuals with SAD and controls. As expected, individuals with SAD displayed greater difficulty when approaching ambiguous faces, than when approaching negative faces. The authors interpreted their findings stating that SAD infered ambiguous faces as threatening, which was in line with previous research showing that these individuals rate non-valenced stimuli as being more threatening when primed by an ambiguous face[75]. Another clinical study employed an AAT design similar to the one used by Heuer and her colleagues[68] in a sample of anxious, depressed patients and controls. In contrasts to what was expected, no consistent associations between the automatic approach-avoidance tendencies and psychiatric measures were found. The authors discuss the

settings of the task to have been suboptimal[76] (relative low number of trials per condition) .

One other study[77] examined approach-avoidance tendencies to images of spiders in subjects with fear of spiders, depressed participants and controls. Compared to depressed participants and controls, spider-fearful subjects showed stronger avoidance tendencies for spider images than for the neutral images[77], further supporting the idea that enhanced avoidance tendencies to disorder-relevant stimuli contribute to pathological behaviour.

In the context of Obsessive Compulsive Disorder (OCD), a study[78] measured approach-avoidance tendencies to contamination-related images (dirty toilets, garbage cans) and neutral environments (household objects) in students with high obsessive-compulsive (HC) symptoms and low obsessive-compulsive (LC) symptoms. As expected, HC participants showed a slower approach of contamination-related than neutral pictures, whereas the LC displayed no differences when pulling both categories of pictures. In particular, the degree of these avoidance tendencies to the contamination-related images was positively associated with stronger obsessive-compulsive symptoms. However, in contrast to the hypothesis, the HC did not push away the negative images faster compared to the LC. The authors discussed whether HC individuals may not necessarily display stronger avoidance tendencies but actually display an impaired inhibition of these tendencies[78].

Taken together, these findings reveal the AAT paradigm to be an efficient tool to investigate pathological approach-avoidance behaviours underlying a variety of psychopathologies, including OCD. In fact, some studies have used a modification of the AAT to alter the automatic tendencies in the context of psychiatric conditions. More precisely, these studies employ a training protocol over a certain period of time whereby subjects also perform joystick movements relative to a disorder-relevant stimuli presented on a computer screen, but opposite to their automatic tendencies[79–83]. This was first studied in the context of heavy drinking, to investigate the possibility to modify increased automatic alcohol approach tendencies, following a training protocol that required subjects to approach alcohol-related images. Results in students followed by alcoholic patients suggested that it was possible to change the initial underlying approach tendencies, which prompted research in other psychiatric conditions. A more detailed explanation of approach-avoidance training studies is provided in the following section.

1.2 Obsessions and Compulsions

1.2.1 Obsessive-Compulsive Disorder

Obsessions and compulsions are the main traits when describing OCD, whereby the former is characterized by persistent and intrusive thoughts that individuals try to ignore, whereas the latter refers to repetitive and ritualistic behaviours that the individual feels "*forced*" to perform. In particular, these excessive compulsions can be performed as a result of an obsession or in a conscious way to avoid an unwanted situational event[84]. In addition, although some individuals are aware that their behaviour is pointless, they feel compelled to perform their compulsions in order to alleviate fear and anxiety, which further increases feelings of being "*unfree*". Thus, obsessions and compulsions not only cause distress but also are time consuming, significantly interfering with a person's normal routine, social activity and relationships[85]. Such perspective is even more concerning given the reports who estimate that the prevalence of people diagnosed with OCD is between 2-3% worldwide[84].

Within the obsession and compulsion symptoms there can be a large variety in terms of their content, meaning that individuals diagnosed with or just displaying OCD traits can differ in the nature of the intrusive thoughts and have different ways in which the compulsions manifest, namely in: fear of contamination towards germs or illness; concerns about safety or making a mistake, leading to checking and/or ordering behaviour; obsessive thoughts of aggression; and difficulty in discarding objects, leading to hoarding behaviour[86]. In fact, these seem to group into six OCD subtypes, namely washing, checking, ordering, obsessing, hoarding and neutralizing[86,87].

With regards to the pathophysiology - although not consistently identified due to the OCD comorbidities, the unknown details of different pharmacological mechanisms of action and consequent behavioural correlations - the disorder has been linked to a dysfunction in the serotoninergic, glutamatergic and dopaminergic systems, particularly in the neurotransmitters' transport and receptor proteins[84,88]. These findings, along with advances in neuroimaging procedures, as mentioned above in section 1.1.4.1 Goal-Directed Behaviours and Habits, allowed to change the focus from anatomical regions to functional networks that encompass different cognitive and behavioural functions in OCD. In particular, the structural and functional changes in specific brain regions in OCD allowed to identify dysfunctional patterns in the so-called cortico-striatal circuits[84,88], including the prefrontal cortex, caudate and putamen. Given the neuroimaging and behavioural findings, several theories based on a pathological imbalance between goal-directed and habitual behaviours have been proposed: dysregulation of neural processes which favours expression of habitual sequences triggered by external stimuli over behaviour flexibility; a deficit in establishing response-outcome/goal-directed behaviours which causes patients to rely excessively on previous

built habits in similar contexts; abnormal regulation of goal-directed and habitual behaviours, by prefrontal cortex subregions (for an extended review see [89,90]).

Additionally, risk factors such genes encoding proteins involved in serotoninergic, glutamatergic and dopaminergic systems, as well as environmental factors such as child trauma can increase the probability to develop the disorder [88,89].

1.2.2 Treatment

Given that OCD is a heterogeneous disorder, as evidenced by its many subtypes together with a spectrum of possible comorbidities such as anxiety[91], tics[92] and other disorders[93,94], treatment intervention in OCD is not straightforward[95] (Abramowitz, Mckay and Taylor 2005). Currently, the main treatment options for OCD include pharmacotherapy and Cognitive Behaviour Therapy (CBT).

1.2.2.1 Cognitive Behaviour Therapy

CBT itself is a combination of two types of therapy, Exposure and Response Prevention Therapy (ExRPT) and Cognitive Therapy (CT), whereby each one targets different domains but complement each other during treatment intervention[96]. In ExRPT, the general protocol consists in repeated presentations of the feared stimulus to the patient. This is performed over the course of multiple sessions, during which patients are asked to prevent engaging in ritualized compulsive behaviours that would, otherwise, attenuate their anxiety[97]. Moreover, the degree of exposure to the feared stimulus is gradual, starting with shorter exposures and progressing to higher distress-demanding contexts so that habituation processes occur: the fear and anxiety previously associated with the feared stimuli is gradually weakened, thereby diminishing the need to engage in compensatory ritualistic behaviours[88]. As such, for the therapy to be successful to patients, throughout the sessions the degree of exposure must be demanding enough to activate feelings of fear, while being low enough to allow habituation processes to develop [98]. Thus, enrolment in ExRPT has an influence on the process of cognitive restructuring, that stems from the habituation to stimuli exposure episodes.

In CT, the aim is to deconstruct the exaggerated and dysfunctional beliefs underlying a patient's symptoms, which then helps them to identify, modify and resist their beliefs[98], including threat estimation, overvalued ideas and perfectionism[99,100]. The sessions are designed to activate and invalidate symptom-related beliefs, with a therapist's assistance[98]. In OCD patients with fear of contamination, CT can "*guide*" patients to conclude that, for example, the thought of not washing hands for three times in a row won't actually result in contracting a severe infection and, consequently, dying. Thus, CT has the potential to influence patients' attitude towards the feared stimulus.

With regards to clinical data obtained in studies that have applied ExPRT in OCD, evidence from meta-analyses have supported its benefits for reducing OCD symptoms[101,102]. In addition, a review on its diverse protocol variations has indicated

that a therapist-supervised ExRPT is associated with a greater decrease in symptoms, which further benefited if it targeted the entire ritualist behaviour and if *in vivo* exposure was followed by imagery, together with an optimal duration of 3 months [101]. As for CT applied to OCD, since it is more recent than ExRPT, studies have evaluated CT's efficacy by also comparing its treatment outcome to the ones obtained by ExRPT. Indeed, a meta-analysis showed that treatment outcomes of CT were not significantly different from ExRPT alone, although both showed benefits compared to all control groups[103]. However, one other meta-analysis showed that - although both groups improved in a follow-up after one year - symptom reduction was faster in ExRPT than CT[104]. These evidences also supported the idea that a combination of these procedures could help in optimizing symptom reduction in OCD, due to their cognitive restructuring and behavioural improvements, respectively, in CT and ExRPT[98]. For additional studies on CT and ExRPT in OCD, in adults and paediatric samples, see review by [96].

1.2.2.2 Pharmacotherapy

The first-line pharmacological treatment option given to OCD patients are Selective Serotonin Reuptake Inhibitors (SSRIs) whose effectiveness in reducing symptoms measured by self-rating scales, at least on the short-term, is evidenced by blinded and placebo-controlled studies[105] (For a detailed review see[106]).

In addition, given the glutamate dysregulation in OCD, as shown by an association between OCD and polymorphisms in the glutamate transporter gene[107], biochemistry abnormalities in glutamate in different brain regions[108] and high levels of glutamate in the cerebral spinal fluid[109,110], among others[106,111], different negative (memantine[112]) and positive (glycine[113]) glutamate modulators have shown to be beneficial for patients. Despite resulting in different ways to target glutamate receptors, these agents seem to affect the neuronal and circuitry level in ways that could be beneficial for long-term treatment[106].

1.*2.2.3 Combination of CBT and Pharmacotherapy*

Considering the brain's physiological processes involved in learning, such as in the acquisition of new behaviours, pharmacological interventions have the potential to enhance the effects of CBT. In particular, given the role of glutamate receptors in modulating the strength of neuronal connections, some studies have shown that pharmacotherapies can be used alongside clinical interventions.

More precisely, in the context of using medication during therapy, a glutamate agonist has been show to accelerate improvements of extinction based interventions, in mice, resulting in facilitated extinction of freezing to contextual cues, and in individuals with clinically significant maladaptive fears, such as anxiety disorders and OCD (for an extended review [114]). In the case of OCD, this could translate into better ExRPT treatment outcome upon confrontation with the feared stimuli. Thus, in theory, a combination of medication with therapy would be more effective in reducing OCD

symptoms. Indeed, a meta-analysis of glutamate receptor agonist has shown that its usage enhances fear extinction during ExRPT in patients with anxiety and OCD, although its effects seemed to decrease throughout CBT sessions[115]. With regards to other types of medication, a double-blinded placebo controlled study comparing CBT, SSRI and their combination found that the effect of ExRPT alone did not differ from ExRPT plus an SSRI, but that the use of both was superior to SSRI alone[116].

1.2.2.4 Treatment Barriers

Despite the treatment options mentioned above, some practical problems arise during implementation of these therapies in a clinical context, specifically in the ExPRT. A review by one study[95] suggested that since the procedures used in CT do not require the prolonged and repeated exposures used in ExPRT, patients are more likely to drop-out of treatment when undergoing the ExRPT compared to the CT, particularly those with strong dysfunctional beliefs. Additionally, there is evidence that 30% of OCD patients are considered treatment-resistant, i.e., fail to respond to treatment including medication, therapy or both, which weighs heavily given the symptoms' profound influence on patients' quality of life and wellbeing[117,118]. In these cases, non-conventional treatment options, reviewed in[119] but not discussed in this thesis, must be considered, such as neurosurgery, electro-convulsive therapy and other pharmacological agents.

1.2.3 Maladaptive Habits in Obsessive-Compulsive Disorder

As previously stated, in individuals with OCD it had been that reported the appearance of intrusive and obsessive thoughts to feared stimuli can trigger avoidance tendencies [84,85]. Viewing OCD from the perspective of a 'behavioral' framework, by using concepts of the Impulsive System theorized in the RIM together with the mechanistic learning processes of S-R strengthening (see section *1.1.1 Reflective Impulsive Model* and *1.1.2 Stimulus-Response Connections* above), the exaggerated impulses to avoid the feared stimuli might originate from maladaptive habits. In particular, these behaviours may stem from a pathological shift of activity from the associate striatum to the sensorimotor striatum. In turn, this could result in a transition from goal-directed to habitual behaviours, the latter being strengthened every time the individual performs ritualist behaviours upon distressing thoughts[90,120,121]. Additionally, an abnormal control by the pre-frontal cortex may contribute to a lack of goal-directed behaviours *versus* habitual responses. Indeed, imaging data has allowed to change focus from specific brain regions to functional networks involving the PFC and basal ganglia, particularly an impaired neurocircuitry termed cortico-striatal-thalamic-cortical loops, which is located amongst frontal-striatal connections[122] (for an in-depth review see [90]).

Thus, evidence so far suggests that processes related to the Impulsive System in the RIM are strongly impaired in OCD. However, the currently available therapies mostly target reflective, conscious processes. As such, developing a therapeutical add-on that

could target such habitual processes directly might provide additional benefit for patients undergoing CBT. Indeed, studies have shown that the AAT could be a potential approach to do just that – target the underlying approach-avoidance tendencies in such a way that they can be modified and normalized.

1.2.4 Approach-Avoidance Training

Due to the importance of this topic in understanding the aim of the current thesis, some important concepts mentioned so far are briefly mentioned now:

The Impulsive System of the RIM contains a bidirectional link between behavioural responses and stimuli evaluation, whereby perception of positive or negative feelings and behavioural responses, separately, can influence an individual's motivational orientation towards stimuli[22]. Moreover, this link seems to be automatic in a way that stimuli valence facilitates compatible responses, such as with arm movements which people perform in response to stimuli in everyday lives[21]. Additionally, it has been suggested that psychopathologies such as addiction and anxiety-related disorders have underlying approach-avoidance motivational imbalances[67], whose measurements can help to further investigate these disorders[123]. The AAT paradigm is a relevant task for this matter because it allows to assess implicit automatic approach-avoidance tendencies, whereby subjects react as fast as possible to stimuli presented on a laptop screen by pushing or pulling a vertical joystick, which makes each image zoom-out or zoom-in, respectively (for a review see[18]). Indeed, some studies have measured automatic approach and avoidance tendencies with the AAT in a variety of psychopathologies, including addiction and anxiety-related disorders (mentioned in section *1.1.6 Approach-Avoidance Measurements in Clinical Psychology and Psychiatry*).

Additionally, taking into account (1) the bidirectional link theory, according to which the behaviour executed in response to a stimulus could itself exert an influence over the stimulus evaluative processes (section *1.1.3 Approach-Avoidance Behaviour*), (2) the evidence that the AAT was able to modify automatic behavioural responses to stimuli in healthy participants (section *1.1.5 Approach-Avoidance Task*), and (3) the motivational approach-avoidance imbalances associated with different psychopathological disorders[67] (section *1.1.6 Approach-Avoidance Measurements in Clinical Psychology and Psychiatry*), this led to the development of AAT training protocols with similar features as described in section *1.1.5.2*, but for psychiatric patients. The purpose of this was to have patients repetitively react to disorder-relevant stimuli in an opposite way of their automatic tendencies, in order to try to modify the valence they attributed towards the stimuli.

The first study to employ such a AAT training protocol investigated the possibility to modify automatic approach tendencies underlying alcohol consumption in heavy drinking students[80]. Following a pre-assessment, subjects either trained to avoid or to approach alcohol-related images. Half of the images used in the assessment were used

in the training, in order to evaluate the generalization capability of the training to images not encountered during training, i.e., that were only seen in the assessments. After training and post-assessment, subjects rated the images and performed a taste test, whereby the amount of alcohol consumed was measured. Results showed that subjects assigned to the avoid-alcohol condition displayed a change from an approach to an avoidance bias towards alcohol-related images, whereby they became faster at pushing away, i.e., avoiding alcohol-related images. Additionally, the training effects generalized to untrained alcohol-related images. More importantly, in subjects who successfully became faster at avoiding alcohol images, these training effects translated into drinking less beer in a taste test after the AAT protocol[80].

Further extending these results, Wiers and his colleagues[79] used a similar AAT design in a clinical sample of alcohol-dependent patients, but with the main difference of four consecutive training days between assessment sessions. Results showed that patients who trained to avoid alcohol-related and to approach non-alcohol images changed their initial approach bias to an avoidance bias towards alcohol-related images. In addition, these training effects appeared to improve abstinence rates one year after the AAT training[124]. In another similar study[81], patients who were undergoing CBT were assigned to either AAT or sham training, in twelve non—consecutive sessions. Results replicated the findings by Wiers and colleagues[79], whereby, in the short-term, the group who received both CBT and AAT training developed avoidance bias for alcohol-related stimuli. In addition, this group reported fewer relapses than the group who received the sham training[81]. In a crucial step to examine if the effect of the AAT training on post-assessment and alcohol consumption was mediated by modified action tendencies or the result of selective attention, a study required students with relatively moderate alcohol consumption to perform a novel task[82]: There, it was assessed how easily participants kept or shifted their attentional focus to a beverage *versus* a neutral image. Additionally, at the end, subjects performed a taste test and filled in questionnaires on their alcohol consumption. Not only did the results replicate the previous findings on inducing avoidance tendencies towards alcohol-related stimuli via an AAT training, but also showed that this was mediated by a change in action tendencies and not selective attention[82]. In contrasts with previous evidence, a study employed a training period of five consecutive days in heavy-drinking students, where they trained either only to avoid alcohol images or to approach and avoid alcohol and non-alcohol images equally[83]. On the last day, subjects participated in a drinking session. Results showed no reduction in alcohol approach tendencies and no decreased alcohol consumption[83].

With regards to SAD, individuals with SA traits tend display reduced approach behaviours, which hamper the development of relationship. In this sense, some studies have examined the effects of training when approaching positive or smiling facial expressions in order to possibly compete with the increased avoidance tendencies. In particular, one study found that after training students showed increased social approach behaviour in a social interaction task with a study volunteer. In particular,

despite there were no changes in the participants' self-reported levels of anxiety after one training session, these self-rated anxiety levels decreased after the friendship-building task[125]. In contrast to this previous investigation, another study did not find group differences in a SA-related outcome explicit measure after a three-session training period, as well as no consistent change in the action tendencies[126].

Concerning AAT training protocols in OCD, one study examined automatic approach-avoidance tendencies in students with high *versus* low fear of contamination[127]. In the single training session, participants were randomly assigned either to the approach condition (mostly approach contamination-related images and mostly avoid neutral images) or to a control condition (approach and avoid both types of images with equal frequency). Upon completion of the computer tasks, subjects rated the images used in the AAT and performed a behavioural approach test. This comprised of three tests, each one with six steps, in order to assess avoidance behaviour to a variety of stimuli such as dirty clothes, insects and dirty toilets. Results showed that participants in the approach condition increased decreased their avoidance tendencies for the contamination-related images, while also completing more steps in the behavioural test[127]. Another recent study also showed positive findings, but in an online intervention. Upon one month of training to approach contamination-related pictures, using a computer mouse to perform the push and pull movements, patients diagnosed with OCD showed a stronger reduction of OCD-related symptoms, compared to patients in a control group However, the study did not assess implicit or explicit measure before and after the training period[128].

1.3 Current Thesis

Previous evidence suggest that AAT training protocols have the potential to modify pre-existing automatic response tendencies, both in healthy individuals (see section *1.1.5 Approach-Avoidance Task*) and in patients with mental disorders (see section *1.1.6 Approach-Avoidance Behaviour in Clinical Psychology and Psychiatry*). Given this evidence, the current thesis aims at extending the work on modifying pathological automatic response tendencies through a computerized AAT training protocols in individuals with OCD-like traits.

As previously mentioned in section *1.2.2 Treatment*, ExRPT and CT are the two main OCD treatment-based therapies whose aims are focused on behavioural habituation and disentanglement of dysfunctional beliefs associated with the feared stimuli, respectively. Despite the evidence in symptom reduction for OCD patients who undergo these therapies their inherent procedures, however, require a significant amount of cognitive effort to execute (see section 1.2.2.4 Treatment Barriers).

Therefore, given the AAT's ability to probe implicit and modify automatic processes (see section 1.1.5 Approach-Avoidance Task), the main research question underlying the current thesis is whether OCD patients benefit from an AAT training protocol as an add-on while they undergo ExRPT, in order to facilitate an implicit acquisition of new and opposite (approach) automatic tendencies for the feared (contamination-related) stimuli. To address this, an AAT training protocol was pre-tested in healthy individuals with OCD-like traits.

Mechanistically, this thesis assumes that the process above would happen through the building-up of new behaviour (approach responses) which would compete with the enhanced avoidance behaviour throughout the task's trials, until the former would become the new automatic response that could prevail even under stress. In more detail, taking into account the bidirectional link theory between behavioural responses and stimuli evaluation, together with the potential of the AAT to modify automatic behavioural responses, this thesis assumes that by repeatedly approaching a feared stimulus through AAT training, the underlying enhanced avoidance automatic tendencies can be contradicted. This contradiction is theorized to happen through the gradual building-up and strengthening of new S-R connections that link the (new) approach response to the feared stimuli, until they gradually become strong enough to compete and contradict the (old) S-R connections, i.e., avoidance responses. More importantly, once the new responses strengthen enough to become habitual responses, they could persist even in situations with depleted cognitive control, such as when confronting contamination-related stimuli either during ExRPT sessions or in daily routine. Moreover, if correctly acquired, the new automatic responses should be displayed as approach biases even when not directly paying attention to the

contamination-related stimuli. Thus, this procedure might be able to reduce patients' difficulty during ExRPT and the risk of drop-outs during its procedure.

To assess the research question mentioned above, , healthy controls with high *versus* low fear of contamination traits were recruited in order to more easily modify their automatic avoidance tendencies due to weaker avoidance tendencies and to avoid comorbidities common in OCD that could interfere with the study aim The healthy controls were pre-selected according to their fear of contamination levels, either with higher fear of contamination traits (higher fear of contamination group; HG) or low fear of contamination traits (low fear of contamination group; LG).

The AAT protocol that subjects underwent consisted mainly of assessment and training sessions. The latter (AAT Training) consisted of five consecutive sessions in separate days, whereby, in each session, all participants were instructed to repeatedly approach contamination-related pictures and avoid neutral pictures as fast as possible, and afterwards rate those images. In addition, participants performed two assessment sessions (AAT assessment), one before and one after training, where all images had to be approached and avoided an equal number of times, as well as explicitly rated. Throughout the protocol, participants performed push or pull movements with a joystick connected to a computer, which made each image progressively zoomed out or zoomed in, reflecting avoid and approach responses, respectively. To test for training effects, several aspects were analysed, mainly: RTs throughout the training period; comparison between pre and post-training assessment RBs and explicit ratings displayed for each stimuli.

1.4 Hypotheses

At the design stage of the current thesis, all hypotheses for each of the AAT versions, i.e., AAT training, AAT assessment and AAT arrow, were established. These were the following:

1.4.1 AAT Arrow

- ✓ Regarding participants' general motor biases when reacting to the directions of neutral arrows (push *versus* pull) without any additional images, it was expected for both groups not to show any approach or avoidance RB.

1.4.2 AAT Training

- ✓ The HG would initially have higher RTs and a slower learning rate throughout the training period when performing the *approach negative* condition, compared to LG, due to the HG's stronger tendencies to avoid negative stimuli. These would translate, respectively, in a higher intercept and a less steep slope in the HG's learning curves, for the *approach negative* condition;

- ✓ Regarding the comparison between conditions within groups, it was anticipated that the LG – in comparison to the HG – would have a more pronounced difference between the two conditions, due to the HG's stronger tendencies to avoid negative stimuli. As a result, this would reflect in a less pronounced decrease of RTs in the *approach negative* condition, compared to the *avoid neutral* condition, throughout the training period in the HG.
- ✓ It was predicted that the LG would experience a stronger reduction of the unpleasant feelings towards the negative images throughout the training period, than the HG, due to the LG's less pronounced sensitivity for the negative images.

1.4.3 AAT Assessment

- ✓ The LG would experience a stronger avoidance RB reduction for the negative images from the 1st to the 2nd assessment, compared to the HG, due to the HG's should have stronger fear of contamination traits that could hamper the training effects.
- ✓ This pattern would be specifically visible for the trained negative images, compared to the untrained images (both with medium strength of content).
- ✓ With regard to the other untrained (assessment-only, weak and strong) negative images, the avoidance RB would decrease the most for the LG in the weak negative images, from the 1st to the 2nd assessment. This hypothesis comes from the fact that approaching negative images whose content is less pronounced would be easier than to approach negative images with more pronounced (negative) content. In addition, this would be specifically observed in the LG due to their less pronounced sensitivity for the negative pictures.
- ✓ Concerning the ratings participants performed before and after the 1st and 2nd assessment sessions, respectively, the same hypotheses were established as the ones for the analyses of RBs, whereby an avoidance RB reduction for the negative images would be complemented by a decrease in the unpleasantness feelings elicited by those same images. Likewise, the same pattern expected for the RBs in the trained *versus* untrained images and in the weak *versus* strong negative images would be observed in the unpleasantness feelings.

2.Methods

Content:

- ✓ ***General Description of the AAT***
- ✓ ***AAT Versions***
- ✓ ***Sample Description***
- ✓ ***Data Pre-Processing***
- ✓ ***Mixed-Effects Models***

2.1 General Description of the AAT

2.1.1 General Procedure

Throughout the period of the AAT protocol, participants sat in front of a desk with a joystick connected to a laptop, with which they had to either approach or avoid the stimuli presented on the laptop screen, according to the instructions provided. Correspondingly, participants had to react *as fast as possible* by either pulling or pushing the joystick with their dominant hand, which would make the stimuli instantaneously, but swiftly, decrease or increase its size, respectively. With this framework, participants were repeatedly trained to approach negative stimuli and avoid neutral stimuli, throughout five sessions on five consecutive days. During the protocol, participants were supervised by an instructor, which provided information about the instructions and time estimation for each AAT version, as well as support for any doubts.

Different neuropsychological questionnaires and different versions of the AAT were applied for distinct purposes, which will be further described below. These were based on a previously established AAT design that used this task to study the underlying mechanisms involved in the acquisition of new habits that contradict automatic tendencies[129]. Additional modifications were implemented to better capture the subjects' emotional valence towards the presented stimuli, to increase the effect of the task training on the subjects' habitual component and to reduce the overall noise in the data. Both, the details of each AAT version and the new adjustments are mentioned in the following sections below.

On the first day of the protocol, participants were given an explanation of the study, regarding the following: Instructions; Aim of the study; The AAT protocol; The possibility to withdraw their consent at any time and; Data protection regarding their personal data, questionnaires and training results. Upon this explanation, subjects were given an informed consent to read and sign. Afterwards, the instructor indicated a general set of instructions for the subjects to follow throughout the AAT protocol, in order to minimize the amount of noise in the data. These were the following: (1) To sit in a comfortable posture, with straight back and no crossed legs, in order to establish a standard sitting position across all participants; (2) To use the dominant hand to hold

Figure 1: Performance of the Approach-Avoidance Task – Every participant was given the same set of instructions to ensure that all performed the entirety of AAT in a standardized manner. Participants' performance was always supervised by an instructor.

the joystick and to place the non-dominant hand on the joystick's base, the latter to provide stability during the movements; (3) To react as fast as possible with the joystick according to the instruction given and with the least amount of errors, since the aim was to analyse the RTs and; (4) To focus on the stimuli presented on the laptop screen. After the instructions were given, participants performed a Pre-train version of the AAT followed by the Arrow version, both of which are explained in the section *2.2 AAT Versions* (below).

2.1.2 Materials Used

The joystick Logitech Extreme 3D Pro was connected to a DESKTOP-386HT0I Toshiba. The former was placed between the individual and the laptop, close to the edge of the desk, so that participants held it without arm extension. With this arrangement, the laptop screen was approximately 60 centimetres away from the seated participants. To avoid discomfort and fatigue in the arm due to the joystick motions, and thus slowness in the RTs, a thin cushion was used to provide support for the elbow to rest on.

All AAT versions were programmed and developed using MATLAB R2012b version with the Psychophysics Toolbox extensions[130,131] which read, processed and saved participants' output throughout each AAT version in single excel files. All graphics and statistical analyses on the participants' behavioural performance, contained in the excel files, were performed with the programming language R[132] version 1.2.1335. In particular, the following packages were used: *lme4* for the Mixed-Effects Models analyses; *multcomp* for the contrasts the post-hoc testing; and *stats* package for model fitting.

Additionally, neuropsychological questionnaires were used to assess participants' general mood and personality traits during the course of the AAT protocol. These were the following: Behavioural Inhibition and Approach System (BIS/BAS) Scale,

Beck Depressive Inventory (BDI), the Symptom Checklist 90 (SCL-90), the Neuroticism-Extraversion-Openness Five-Factor Inventory (NEO-FFI), the White Bear Suppression Inventory (WBSI) and the Positive and Negative Affect Schedule (PANAS). These are further explained in section *2.2.3 Questionnaires*.

2.1.3 Stimuli

The stimuli used in the task comprised of 32 images organized in four categories (8 images per category): positive, negative and two distinct neutral sets, neutral-street and neutral-kitchen. The positive images displayed beach-related contents in order to try to elicit feelings of enjoyment and pleasure, while the negative images displayed dirty toilets in order to elicit feelings of disgust. The two categories of neutral images were used as control conditions for the negative and positive images. More precisely, a set of neutral-street-related images was used as the counterpart for the positive images, while the set of neutral-kitchen-related images was used as the counterpart for the negative images.

The images of the negative and neutral-kitchen categories were chosen upon analysis of the online questionnaire (see *Annexes*). As for the positive and neutral-street images, the same four images of each category used in a previous study[129] were used in the current thesis, simply adding in new four images to each category to ensure that all four categories contained an equal number of images, i.e., eight images per category. The newly added images were selected through the Common Objects in Context database[133] which meet the following criteria: sunny weather, clear water with no turbulence, no presence of people and animals, for the positive category; clear sky, no walking pedestrians, no traffic jam and no red colour-displaying structures (including fire hydrants, stop signs and traffic lights), for the neutral-street category. In case of additional editing to remove traces of unwanted elements, the images were edited with the photo editor software *Photoshop*. Afterwards, all images were resized to a dimension of 400x300 pixels.

In all AAT versions, the images were presented at the centre of the screen, upon which the image would change its size according to the movement of the joystick. A vector of 500 equally spaced image sizes was created for each image, so that at the beginning of each trial, at point 250, the image would appear at the centre of the screen in 400x300 dimension. Movement of the joystick would make the image instantaneously increase or decrease its size, according to the direction of the movement, in a rapid and swiftly manner. Once the joystick was at either end, and, thereby, the image at its maximum or minimum size, the image remained on the screen for 1 second, before the next trial started. This time added per trial will now be referred to as "*lock time*". This modification was based on a previous study that applied the same feature in an AAT design for OCD patients[128].

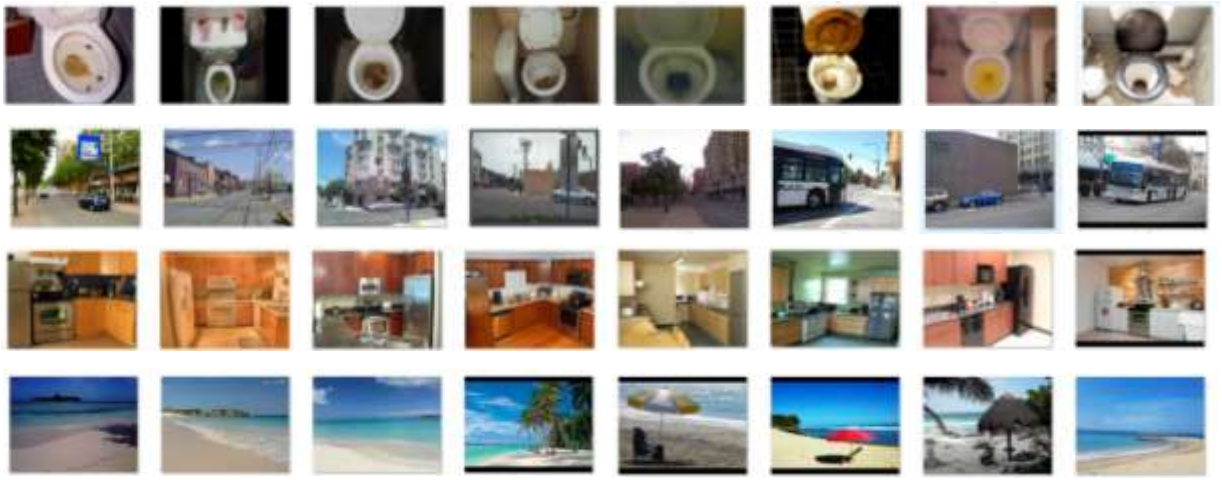


Figure 2: AAT Images - Images used in the AAT, arranged by their content characteristics. The images in the first row depict dirty toilets (contamination-related content), which were used as negative stimuli. The images in the second row depict city street-related environments, while the images in the third row depict kitchen environments. Both of these were used as neutral stimuli. The images in the fourth row depict beach-related scenarios, which were used as positive stimuli.

2.1.4 Brief Presentation of Images

On the first day of the protocol, after the AAT Arrow version and before the 1st AAT Assessment, participants were shown each of the 32 images for 5 seconds, on the laptop screen. Initially, a message appeared on the screen saying "*Observa livremente as seguintes imagens. Pressiona a tecla A para começar.*" ("*Freely observe the following images. Press A to start*"), whereupon the first image was displayed for 5 seconds, followed by a white screen with a black dot at the centre for 1 second, followed, in its turn, by the next image and so on. The 32 images were shown in the following order of categories: neutral-kitchen, neutral-street, negative and positive. This was done in order to prevent (positive or negative) contamination of the neutral categories. In addition, each set of images was shuffled so that the order of image presentation within each category was randomized for each participant.

2.2 AAT Versions

2.2.1 AAT Protocol

To ensure that all participants performed the AAT under the same conditions, a standardized protocol, which contained the questionnaires and AAT versions to be performed every day by every subject, was improved based upon a previous study that applied the AAT training in healthy controls[129]. This improved protocol is shown in the figure below. Along with the AAT performance, additional questionnaires were filled-in in order to evaluate participants' general mood and personality traits that might have influenced their performance in the experimental task.

<table>
<tr><th>1st Day</th><th>2nd Day</th><th>3rd Day</th><th>4th Day</th><th>5th Day</th></tr>
<tr><td>Pre-Test
AAT Arrow
Image Presentation</td><td rowspan="2">AAT Training</td><td rowspan="2">AAT Training</td><td rowspan="2">AAT Training</td><td rowspan="2">AAT Training</td></tr>
<tr><td>Ratings & AAT Assessment</td><td rowspan="2"></td><td rowspan="2"></td><td rowspan="2"></td><td>AAT Assessment & Ratings</td></tr>
<tr><td>AAT Training</td><td></td><td></td><td></td><td>Practical Test</td></tr>
</table>

Figure 3: AAT Protocol - Versions of the AAT to be performed by each participant throughout the consecutive five-day period.

2.2.1.1 Pre-Test

The purpose of this version was solely to let participants feel comfortable with both the joystick's shape and motion together with the instructions that were provided, in order to avoid the confounding factor of grabbing and moving a joystick for the first time. Here, participants saw a random noise rectangle with an arrow, whose direction indicated the action to be performed with the joystick, approach or avoid. Each participant performed this version until they felt comfortable when performing the push or pull motions as well as the other instructions regarding posture (mentioned in the section *2.1.1 General Routine*).

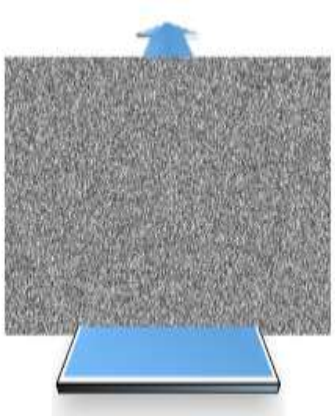


Figure 4: Pre-Test AAT Version - In each trial, participants had to push or pull the joystick as fast as possible, according to the direction indicated by the arrow. This task was performed until each participant felt comfortable with the joystick as explicitly reported to the instructor.

Additionally, participants were explained that depending on their performance, each trial could finish in one of three ways: (1) When mistakenly executing the wrong instruction, such as performing a push motion when the arrow instructed a pull motion or vice-versa, a message would appear on the screen saying "*Errado*!" ("*Wrong*!"); (2) If the full joystick movement time was equal or higher than 1.5 seconds (explained in the subsection *Data Pre-Processing* below), a message would appear on the screen saying "*Mais rápido, por favor*" ("*Faster, please*") and the trial would be repeated; (3) If the action to be performed was correct and faster than 1.5 seconds, the trial would be signalled as correctly performed and it would move on to next trial. All other AAT versions also had these features.

2.2.1.2 Arrow version

The arrow version was designed to check if participants had any general reaction tendencies when performing either the pushing or pulling motions with the joystick, without any affective stimulus. More precisely, with the RT data acquired in this version the aim was to identify whether participants had a general preference for performing approach or avoidance responses, which would be reflected in faster RTs in either of those responses. In this version, participants saw a black rectangle together with an arrow whose direction informed the action to be performed, i.e., to either pull or push the joystick. The push and pull motions were performed for 10 trials each, randomized throughout 20 trials.

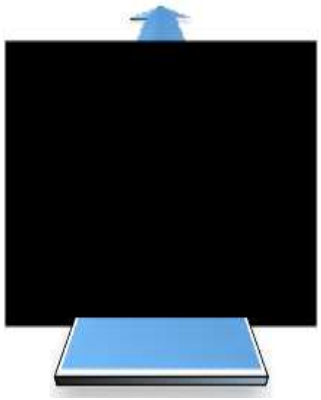


Figure 5: Arrow AAT version - In a similar fashion to the Pre-Test, participants had to push or pull the joystick as fast as possible, according to the direction indicated by the arrow.

2.2.1.3 Training version

Participants were instructed to perform two conditions, the *approach negative* condition and the *avoid neutral-kitchen* condition throughout five training sessions. The purpose of this AAT version was to make participants react in an opposite way to the automatic avoidance tendencies they had towards the negative images. The *avoid neutral-kitchen* condition was used as control condition relative to the *approach negative* condition.

Figure 6:Instructions at the beginning of the AAT Training Version - Each participant had to approach, i.e., pull the joystick closer when the screen displayed one of the two negative images ("Quando vires estas imagens, aproxima-as de ti puxando o joystick"; "When seeing these images, pull them closer to you with the joystick"). Inversely, they had to avoid, i.e., push the joystick away when the screen displayed one of the two neutral-kitchen images ("Quando vires estas imagens, afasta-as de ti empurrando o joystick"; "When seeing these images, push them away from you with the joystick").

For each training session, each condition had two images, whereby each image was used for 30 trials. In other words, each negative image was approached 30 times and each neutral image was avoided 30 times, yielding 60 trials per condition and 120 trials in total per training session. The *approach negative* and *avoid neutral-kitchen* trials were randomized throughout the 120 trials. The RTs to complete each trial were obtained. After performing all 120 trials, participants rated the same images that were shown during the training, in terms of their pleasantness.

Figure 7: AAT Training and Ratings – During the AAT training, all negative images had to be avoided and all neutral-kitchen had to be approached, with a pull and push joystick motion, respectively (left side of the figure). After each training session, all images were rated through a horizontal bar with two ends, coded -100 in the far-left ("Unpleasant") and 100 in the far-right ("Pleasant"). Here, each image was rated based on its "pleasant"/"unpleasant" characteristics, whereby the question "O quão desagradável ou agradável é esta imagem para ti, neste momento? ("How unpleasant or pleasant in this image for you, at this moment?") was presented on top of the horizontal bar (right side of the figure). Participants used a computer mouse to select a point anywhere between the two ends of the horizontal bar.

Each participant had a different and predetermined set of two negative and two neutral-kitchen images that he/she trained with in the *approach negative* and *avoid neutral-kitchen* conditions, respectively. This predetermined assignment was based on a list previously made with all the possible combinations of two images between the four

medium content images (see *Annexes*). More precisely, based on the analysis of the online questionnaire, four medium content negative images and four neutral-kitchen were chosen for the AAT Training. Since the aim was for each participant to train with two images for each condition, there were 6 possible pairs in each condition that could be assigned to each subject (from the 4 images content pick 2, without repetitions) and a total of 36 combinations (6 pairs for the negative and 6 for the neutral), when taking into account the different pairs for the negative and neutral-kitchen images. Thus, each participant was assigned two medium content negative images for the *approach negative* condition and two neutral-kitchen images for the *avoid neutral kitchen* condition.

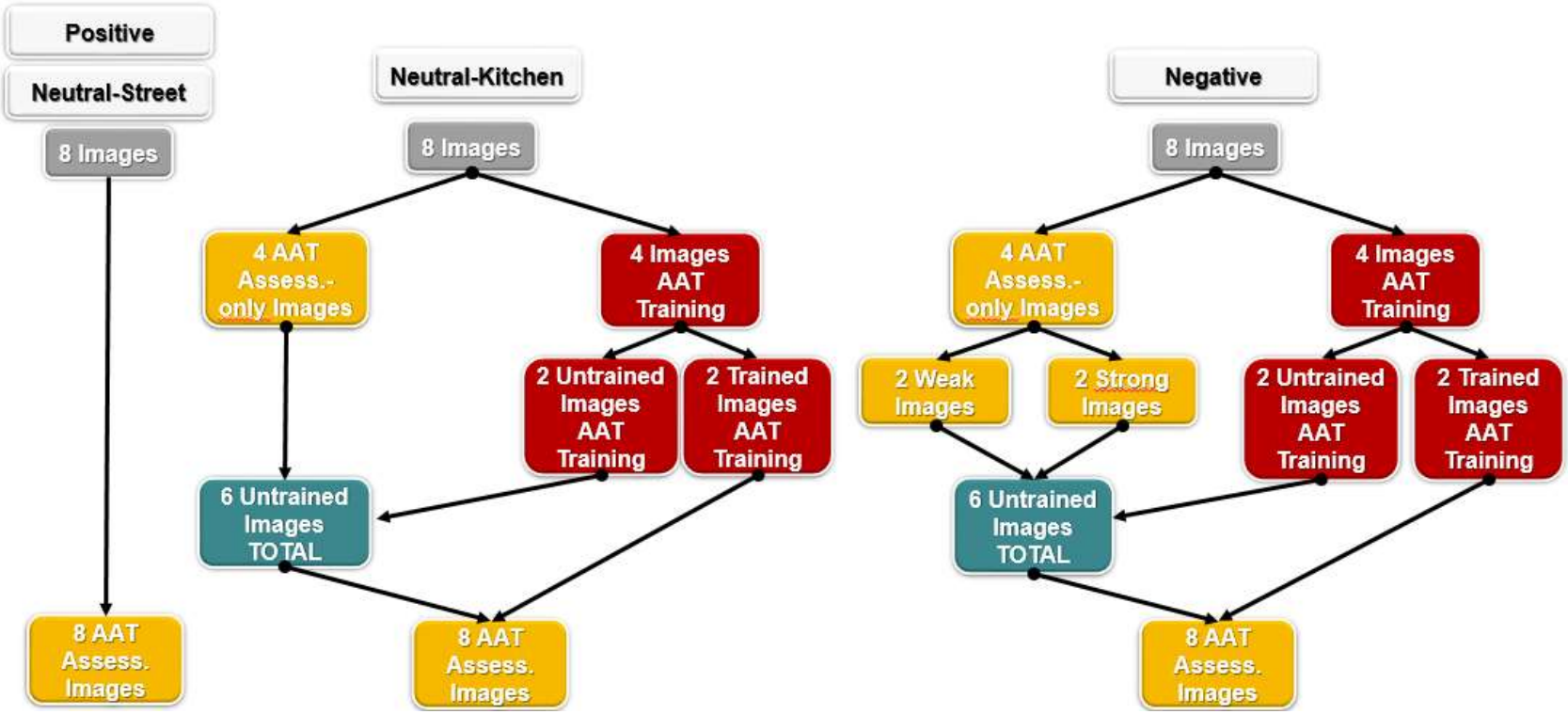


Figure 8: Schematics of the Images used in all AAT Versions - All eight images (grey boxes) of each category (white boxes) were shown in the AAT assessment (lower yellow boxes). However, in the negative and neutral-kitchen categories, half of the images used in the AAT assessment, with equally medium content strength, were allocated to be also used in the AAT training (red boxes). More importantly, in a pseudo-randomized order, in each subject half of these training-allocated images were shown in the AAT training while the others were shown only in the AAT assessment (red boxes split). The other half of the negative and neutral-kitchen images was shown only in the AAT assessment (middle yellow boxes). In particular, in the negative category, these assessment-only images consisted of two images with weak content and two images with strong content characteristics.

2.2.1.4 Assessment version

This version was used to evaluate participants' automatic RBs towards the content displayed in the negative, positive, neutral-kitchen or neutral-street stimuli, before and after the training period. To do so, each image was presented together with an arrow whose direction indicated the action to be performed (approach or avoid) and for which participants were instructed to only pay attention to, rather than to the content of the stimuli. With the RTs acquired for each trial, the RBs that participants displayed when responding to the stimuli were obtained (see section *Reaction Biases Calculation* below). Briefly, faster RTs when performing an avoidance action in response to negative stimuli, compared to when approaching, would indicate an avoidance RB for the negative stimuli. Conversely, faster RTs when performing an approach action in response to positive stimuli, compared to when avoiding, would indicate an approach RB for the positive stimuli.

Each image was approached and avoided 6 times, yielding a total of 384 trials (32 images x 2 actions x 6 repetitions) per assessment session. In particular, an algorithm was used to restrict the way in which the trials presented in the assessment were shuffled, according to the following criteria: (1) there could not be two or more consecutive negative trials for the participant; (2) for every six non-consecutive trials that displayed a negative image, the preceding trial had one of possible six combinations: approach positive, avoid positive, approach neutral-kitchen, avoid neutral-kitchen, approach neutral-street or avoid neutral-street (basically, approach or avoid one random image from each category, except for the negative). The order in which these combinations were placed before every six negative trials was created randomly, and its iteration across the rest of the 384 trials kept changing randomly. Moreover, this randomly created order was different for each assessment AAT version session and for each subject. The role of these algorithmic shuffle restrictions was to help normalize the contamination effect that each previous trial could have over the following negative trial in terms of emotional bias and, therefore, on the RTs displayed.

Figure 9: AAT Assessment and Ratings - All images from the four categories, i.e., negative, positive, neutral-kitchen and neutral-street, were approached (joystick pulling motion) and avoided (joystick pushing motion) an equal number of times. Participants were instructed to perform the action as fast as possible while paying attention to the direction of the arrow, in order to measure participants' automatic reaction bias towards the stimuli (left side of the figure). The ratings were performed in an equal manner as in the AAT training, via a horizontal bar with two ends (right side of the figure).

In addition, participants rated all 32 images in terms of their pleasantness ("*O quão desagradável ou agradável é esta imagem para ti, neste momento?*"; "*How unpleasant or pleasant in this image, right now?*"), reaction elicited ("*A minha reação a esta imagem é...*"; "*My reaction to this image is...*") and comprehensiveness ("*Quão fácil/difícil é perceber o conteúdo da imagem?*"; "*How easy/hard is it to understand the content of the image?*"). These ratings were performed through an identical framework as in the AAT Training, i.e., through a computer mouse to select a point anywhere between the two of a horizontal bar displaying "*Desagradável"/"Agradável"* ("*Unpleasant"/ "Pleasant"*), "*Afastar"/"Aproximar*" ("*Avoid"/"Approach*") and "*Difícil"/"Fácil"* ("*Hard"/"Easy"*), for each question, respectively. Depending on whether it was the 1st or 2nd assessment session, these ratings were performed before or after the assessment AAT. Each AAT assessment session was split in two halves, whereupon *MATLAB* restarted in order to avoid participants' arm to fatigue and to avoid the presentation programme *MATLAB* to overload the memory capacities of the laptop, slow-down and possibly crash during the task.

2.2.1.5 Practical Test

To further measure the potential effect of the AAT Training on participants' habitual behaviour, i.e., to correlate the effects of the training with stimuli related to real-life context, an additional final test was performed at the end of the 2nd AAT assessment, on the last day of the protocol. First, it was measured the time it took for each participant to pull a chair and sit down on it in front of the laptop, whereby the pillow to sit on displayed an image of a dirty toilet. With the purpose of avoiding confounds related to sitting on the pillow for the first time, throughout the whole week participants sat on a chair with a pillow in a white cover. Solely for this test, the pillow's cover was purposely changed, unbeknownst to each participant who was filling questionnaires in another chair.

Figure 10: Pillow with Modified Cover- Contamination-related image depicted on the pillow cover used in the practical test.

Upon asking participants to sit down on the same pillow now with the modified cover, as soon as the chair was pulled away from the table, the instructor pressed the spacebar key on the laptop to start recording the time it took participants to sit and press the same key on the laptop. During this time, the laptop screen displayed the instruction "*Depois de te sentares pressiona a barra de espaço*" ("*After sitting down, press the spacebar*").

Afterwards, participants rated novel images from three categories, positive, neutral and negative. In these ratings, participants had to rate the same set of images as the ones used in a previous study[129], in terms of their pleasantness. More importantly, these images were distinct from the ones used during the AAT protocol, as they were meant to display more general, everyday stimuli.

Figure 11: Practical Test Images - Novel negative, positive and neutral images used in the ratings after the sitting test. These images were taken from the International Affective Picture System[134] (IAPS) database and from Microsoft's Common Object in Context[133] (COCO) database. Their respective assigned numbers in IAPS are the following: 4689 and 8501, for the negative images; 1280, 1525 and 6250, for the neutral images; 7010, 7175 and 7090, for the positive images.

2.2.2 Summary of Modifications

Several modifications compared to a previous study[129] were implemented in the AAT protocol and in its underlying *MATLAB* code. These modifications have already been mentioned throughout the previous sections and are summarized here:

- A brief presentation of all 32 images before the 1st assessment AAT version, on the first day of the Protocol (see section *2.1.4 Brief Presentation of Images*).
- Restrictions in the way how all negative trials were shuffled, in the AAT assessment version (see section *2.2.1.4 Assessment Version*).
- A lock time at the end of each trial, whereby the stimuli remained on screen for 1 second (see section *2.1.3 Stimuli*).

2.2.3 Questionnaires

As previously mentioned, neuropsychological self-report questionnaires were used during the protocol to assess participants' general mood. As their data was not analysed in the current thesis, only a brief explanation is provided for each one. The BDI measures the presence and degree of depression, with items covering various symptoms including sadness, pessimism, suicidal thoughts, loss of interest, irritability, among others[135] (performed at the end of day 3). The SCL-90 provides scores on five symptom dimensions, which include somatization, obsessive-compulsive, interpersonal sensitivity, anxiety, and depression[136] (performed at the end of day 3). The BIS/BAS scale measures individuals' motivation to avoid aversive outcomes and to approach beneficial outcomes[137] (performed at the day end of day 5). The NEO-FFI measures hostility, depression, self-consciousness, impulsiveness and anxiety[138] (performed at the end of day 5). The WBSI measures thought suppression, which is related to depression, anxiety and to obsessive-related thinking[139] (performed at the end of day 5). The PANAS is a mood scale designed to provide information about positive and negative affect at the current moment[140] (performed at beginning of each day, and before and after the practical test). Analyses of these questionnaires is out of scope for the current thesis. Future analyses will into the data contained in them.

The translation procedures implemented to obtain the Portuguese versions of the BSI and OCI-R scales used in this thesis were performed by different personnel in the laboratory, based upon the translation guidelines established by Harkness & Schoua-Glusberg[141]. Briefly, these guidelines assure minimal semantic and grammatical loss during translation, within the constraints of what is possible for a given context[141]. (this framework has been applied, for instance, to transnational social sciences research across Europe*). According to these guidelines, the translation process is composed of independent translations made by native speakers followed by a review that combines

The European Social Survey (ESS), in short, is an "…academically-driven cross national survey…*" that "*…measures the attitudes, beliefs and behaviour of diverse populations…*" every two years, in the context of social sciences research (https://www.europeansocialsurvey.org/about/). To do so, the ESS uses specific procedures to ensure an effective translation of the selected questionnaires to each country language's, in order to counteract the subjective nature of translation procedures.

all versions and afterwards by an adjudication, whereupon the final translating decisions are made. Furthermore, all of these steps are documented in the form of alternatives, uncertainties and comments (for instance, in a Word document), in order to better decide the final translating decisions[142].

For the BSI, the Portuguese translation was performed due to experience obtained in a previous study in the laboratory (unpublished): Feedback was given to the instructors, in particular that some items were hard to understand / interpret the existing Portuguese version[143]. Thus, to obtain the BSI Portuguese version used in the current work, a Portuguese translation of the English version[144] was done in the laboratory according to the guidelines described above. A reviewer compared this version to the existing version[143]. Afterwards, an adjudicator combined the two versions and performed the final translating decisions.

As for the OCI-R, a similar procedure was used: A reviewer analysed the pre-existing OCI-R Brazilian-Portuguese version[145] and the Portuguese version that had been previously translated in the laboratory in 2015 for a master thesis[129]. Likewise, an adjudicator combined all versions and performed the final translation decisions. For this process, the pre-existing OCI-R Portuguese version that was performed by Cardoso *et al.* in a master thesis in 2015[146], available online in 2016[147], and published in a scientific journal in 2017[148], was not taken into account for the current work. This is because the current thesis is part of a bigger project which started in 2015 (Reference of the Ethics Committee of the Medical Academic Centre in Lisbon: 251/15), a time at which Ismael's work was not yet publicly available. A brief analysis to compare of the quality of the OCI-R and BSI questionnaires used in the current thesis with the respective validated Portuguese and the original English versions was performed (see section *6.4 Psychometric Analysis* in the Attachments), which indicated that quality of the OCI-R and BSI versions used here were comparable to the previous versions.

With regards to the Portuguese versions of the other questionnaires whose data were not analysed in this thesis, two of them had already been translated in the laboratory for previous projects (including the BIS/BAS and Edinburgh questionnaire). As for the rest of the questionnaires, the existing Portuguese versions were used (including the BDI-II[149], the NEO-FFI[150], the PANAS[151], the SCL-90[152] and the WBSI[183]).

2.2.4 Data Collected Throughout the Protocol

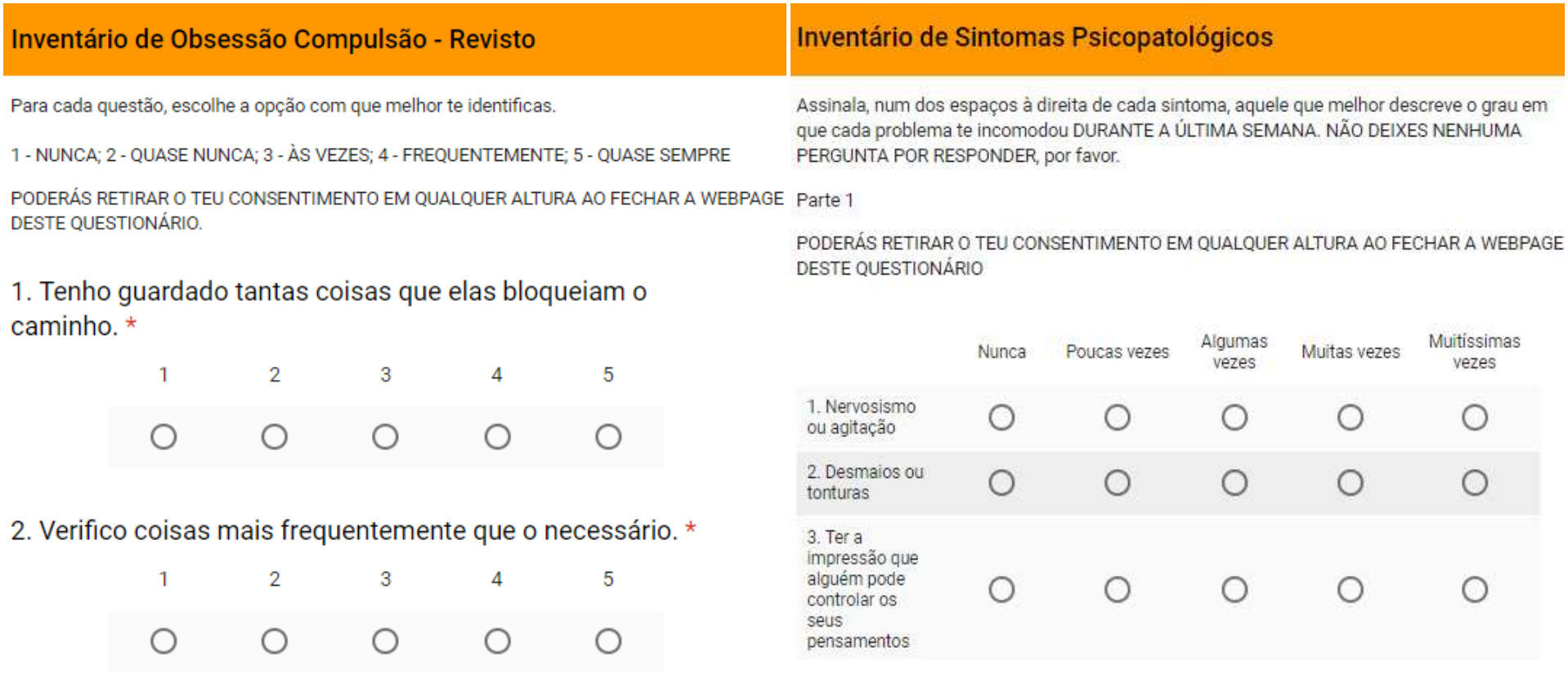
Each session of the protocol provided us with the following outputs: (1) the initiation times in milliseconds (ms), i.e., the time it took to tilt the joystick in 20 degree-angle relative to the initial resting position, once each image appeared on the screen, in each trial; (2) the full joystick movement in ms, i.e., the time it took to completely pull or push the joystick until it reached the end position, in each trial; (3) a time-position vector of the joystick for each trial, which displayed the position of the joystick every 50 ms, relative to the joystick's initial resting position; (4) the score ranging from -100 to 100, given by each subject when rating each image with regard to each question.

2.3 Sample Description

At the initial stages of the current thesis, an estimation of the average number of individuals with higher fear of contamination traits in a student population was performed, to know beforehand the number of students that would be needed to screen. Since the intention was to collect 25 participants with high fear of contamination traits, besides the other 25 participants with lower traits for which it was theorized to be more common, a brief analysis indicated that 250 students would have to be screened to get 25 students with high fear of contamination. For more details on how this estimation was performed, see the *Annexes* chapter.

A standard email containing a brief description of the current study was sent to the various faculties within the University of Lisbon. The email stated that healthy individuals were needed to investigate the attentional and emotional effects of a computer game, the latter having the potential to facilitate therapy in OCD patients in the future. The email also contained a link to an online webpage in google forms, for participants to fill in the Portuguese versions of the Obsessive Compulsive Inventory-Revised[153] (OCI-R) and the Brief Symptom Inventory[144] (BSI) scales. At the beginning of this online webpage, a description of the current study stated the instructions, the aim of the study, an overview of the AAT protocol, possibility to withdraw their consent at any time and data protection with regards to their personal data. After this, the exclusion criterion was stated, for which participant had to read before the filling-in OCI-R and BSI scales, which included "*I do not have a diagnostic or history of psychiatric or neurological disorders, as well as any type of chronic disorders*".

The Portuguese translation for both scales was performed by two independent persons of our laboratory, followed by a group of supervisors who made the final decisions. Depending on individuals' scores in the two scales, those who meet the requirements were sent a follow-up mail to schedule their participation. The requirements are explained now below.



Inventário de Obsessão Compulsão - Revisto

Para cada questão, escolhe a opção com que melhor te identificas.

1 - NUNCA; 2 - QUASE NUNCA; 3 - ÀS VEZES; 4 - FREQUENTEMENTE; 5 - QUASE SEMPRE

PODERÁS RETIRAR O TEU CONSENTIMENTO EM QUALQUER ALTURA AO FECHAR A WEBPAGE DESTE QUESTIONÁRIO.

1. Tenho guardado tantas coisas que elas bloqueiam o caminho. *

1	2	3	4	5
○	○	○	○	○

2. Verifico coisas mais frequentemente que o necessário. *

1	2	3	4	5
○	○	○	○	○

Inventário de Sintomas Psicopatológicos

Assinala, num dos espaços à direita de cada sintoma, aquele que melhor descreve o grau em que cada problema te incomodou DURANTE A ÚLTIMA SEMANA. NÃO DEIXES NENHUMA PERGUNTA POR RESPONDER, por favor.

Parte 1

PODERÁS RETIRAR O TEU CONSENTIMENTO EM QUALQUER ALTURA AO FECHAR A WEBPAGE DESTE QUESTIONÁRIO

	Nunca	Poucas vezes	Algumas vezes	Muitas vezes	Muitíssimas vezes
1. Nervosismo ou agitação	○	○	○	○	○
2. Desmaios ou tonturas	○	○	○	○	○
3. Ter a impressão que alguém pode controlar os seus pensamentos	○	○	○	○	○

Figure 12: Questionnaires used in the Online Questionnaire- All of the items from the Obsessive Compulsive Inventory-Revised (OCI-R) and Brief Symptom Inventory (BSI) were included in the webpage of the google form. There, participants would rate from 1 (“Nunca”; “Never”) to 5 (“Sempre”; “Always”) in the OCI-R scale (left panel) and “Nunca” ("Never”) to “Muitíssimas Vezes” (“Very Frequently”) in the BSI scale (right panel), respectively. The score each participant obtained for each scale was afterwards calculated based on their answers.

In short, the 18-item OCI-R scale provides scores on six subscales, corresponding to the six obsessive-compulsive subtypes that are commonly described for OCD. This scale has been shown to have good internal consistency (the ability of specific items to represent each symptom), convergent validity (high correlations with other measures of OCD) and test-retest reliability (stability of the measurements across time)[86]. In the current thesis, the washing subscale of the OCI-R (which assesses washing/contamination concerns relative to germs, dirt, animals or insects) was used to pre-select individuals with higher *versus* low fear of contamination traits, through an established cut-off. In particular, this cut-off was based on previous studies that applied the OCI-R in OCD patients, showing that the mean score of the washing subscale is more or less equal to 4[86,153]. With this evidence, individuals who scored 4 or higher in the current thesis were labelled as being in the "*high fear of contamination trait group*"(HG), whereas those who scored lower than 4 were labelled as being in the "*low fear of contamination trait group*"(LG).

The BSI is a 53-item symptom inventory designed to assess the psychological symptom status of patients and healthy individuals. It measures nine primary symptom dimensions, namely somatization, obsessive-compulsive traits, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation and psychoticism[144]. In the current thesis, the BSI was used to make sure that all participants were healthy. More precisely, it was used to assess individuals' mental well-being and exclude the ones who scored above the cut-off of 1.7[144], a score which previous research has shown to indicate a higher-than-average level of mental distress.

The total number of students that filled-in the online google forms, which contained the OCI-R and the BSI, was 343. Of these, 80 had a washing subscale score above or equal to the cut-off, 239 had a washing subscale score below the cut-off, and 24 had a BSI score above or equal to the cut-off. Students were then contacted and invited by email, having in mind that the aim was to collect 25 participants with high and 25 participants with low fear of contamination traits, with similar number of females and males in each group.

The final sample of participants recruited consisted of 9 males and 36 females distributed across both groups, yielding a total of 45 participants, 24 in the LG and 21 in the HG. Chi-squared test revealed no significant differences in gender between groups (LG: Males = 6, Females = 15; HG: Males = 3, Females = 18; X-squared = 0.57, p = 0.45). The mean and standard deviation in OCI-R score in the LG was $\overline{X_{OCI-R}} = 13.3$ and $\partial_{OCI-R} = 4.3$, while for the HG it was $\overline{X_{OCI-R}} = 26.3$ and $\partial_{OCI-R} = 6.6$. As for the washing subscale score, the mean and standard deviation in the LG was $\overline{X_{Wash.}} = 0.91$ and $\partial_{OCI-R} = 0.66$, while for the HG it was $\overline{X_{Wash.}} = 5.29$ and $\partial_{OCI-R} = 1.17$. Statistical analysis with T-tests revealed the groups to significantly differ with regard to the OCI-R (t = -6.04, p < 0.001) and Washing (t = -11.70, p < 0.001) scores. The final sample was in the statistical analysis is mentioned in the following section *2.4 Data Pre-Processing*.

Once in the lab, participants signed an informed consent, which had been pre viously approved by the Ethics Committee of the Academic Centre of Medicine of Lisbon, at the beginning of the current thesis.

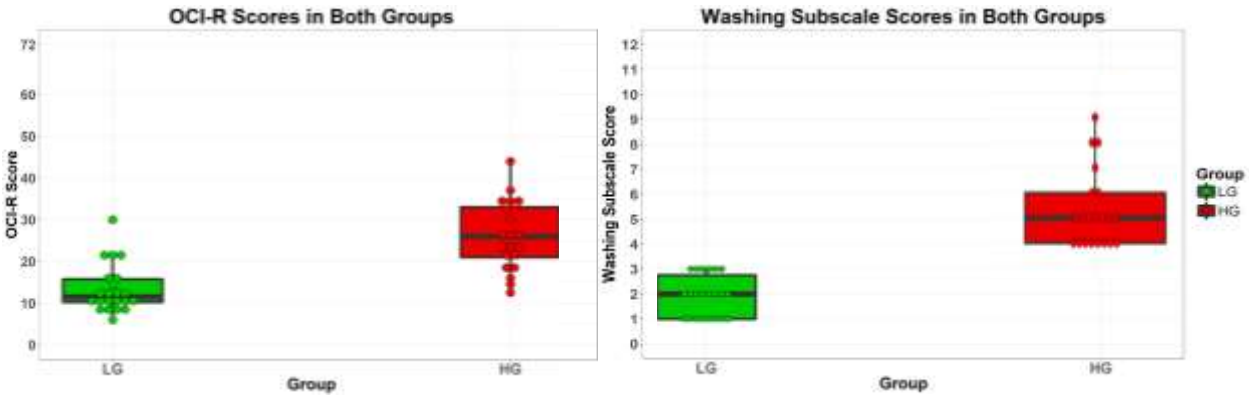


Figure 13: OCI-R and Washing Scores – Visualization of the OCI-R scales (upper panel) and the Washing Subscales (lower panel) scores, in the group with high contamination-related traits (green boxplot) and in the group with low contamination-related traits (red boxplot). The dots inside each box plot represent the scores of subjects for each measure.

2.4 Data Pre-Processing

As a first step, the RTs data contained in the excel files were visually inspected to briefly assess its quality, particularly for any systematic errors throughout the sessions. Afterwards, participant's data were processed in a trial by trial analyses to exclude error-trials, i.e., trials that were considered as either mistakes, outliers, slow relative to the time limit or trials in which the direction of the response had been correct during the movement. The criteria for each of these error-trial types were the following: *Mistakes* – when the action performed was different than the instruction; *Outliers* – criteria based on a previous study[129] (explained below), whereupon RTs lower than 200 ms and higher than three times the interquartile range above the third quartile were considered outlier RTs; *Slow* - when the RTs were higher or equal to 1.5 seconds, whereupon the trial was automatically repeated until performed below the time limit. This value was established in order to impose a time restriction for each trial to be completed within a reasonable time**;** *Corrected trials* - trials at which participants' initial joystick movement was different than the action instructed, but still managed to correct during the trial. With these criteria, the number of error-trials per participants was calculated and excluded from each participants' RTs dataset. Is it also important to refer that the percentage of excluded trials per subject due the above mentioned errors was low, with the LG displaying a maximum of 4.9% and an average of 1.7% excluded trials, and the HG displaying a maximum of 4.7% and an average of 2.2% of excluded trials in the AAT assessment version overall. Therefore, the number of remaining trials for every participant was still large enough to proceed with the statistical analysis.

Regarding the specifics of the outlier criteria, previous literature has suggested two main types of outliers in RTs datasets, namely the short and long RTs. Whereas in the case of the current work, the short RTs were less frequent, due to a natural speed limit for physiological processes that occur during stimulus perception[154], the long RTs appeared more frequently. To process the long RTs, a method was chosen based on a previous study[129], that took into account different approaches and the respective limitations for processing RT distributions. As such, the method of a cut-off at three time the interquartile range (i.e., above the 3rd quartile) was chosen based on its ability to prevent eliminating meaningful information, compared to cut-offs with the mean and standard deviation[155]. Moreover, since the RTs do not follow a normal distribution, a cut-off based on the mean and standard deviation is not the ideal option to exclude observations in the upper tail[154,155], strengthening the choice for using the median and interquartile range criteria.

Thus, nearly all participants fulfilled the requirements regarding data quality, except for one participant: this was due to a coding error that caused abnormally high RTs values. A second participant was also excluded due to a chronic disease, which was one of the exclusion criteria in the pre-selection, but had only been reported by this participant during the training sessions. Thus, the final sample consisted on 41 participants: 20 participants in the LG and 21 participants in the HG.

2.5 Mixed-Effects Models

To perform statistical analyses on the RTs and ratings acquired throughout the AAT protocol, Mixed-Effects Models (MEMs) were applied. The procedure of MEMs is similar to a standard Analysis of Variance (ANOVA), as it tests whether there is a significant difference between independent variables among three or more groups, providing the F-statistic p-values to reject or not the null hypothesis. However, one advantage of the MEMs is that they capture dependency and variability of non-interest variables via random effects parameters that are then incorporated in the models together with fixed effects parameters (hence the name *mixed*), which capture variability of the variables of interest. Additionally, whereas the ANOVA sets equal variance values to all variables, the MEMs allows to "*customize*" the variance, i.e., the design of the model allows one to determine different variances of the variables of interest and non-interest. For a more detail description of the characteristics of MEMs, see [156]).

With regards to the statistical analyses, MEMs are interpreted and tested in the same way as the independent variables in a standard ANOVA (for example, with planned contrasts). As such, when specifying random effects in the MEMs, the influence of random noise is minimized to obtain a more powerful estimate of the fixed effects of interest[156]. For the current thesis, this approach allowed to fit the behavioural data (acquired in the AAT) and to model nonlinear individual characteristics, while also taking into account random and distinct data patterns at two levels. More precisely, these random patterns usually emerge at the subject-level because the RTs are highly variable at the intra-subject level, but still more similar within one person than across subjects. The second level to be captured by the MEMs were the expected influences of the experimental variables (category, content strength, trained vs untrained, session, group). Different MEMs were devised for each AAT version, according to its design characteristics. To analyse the RTs using the MEMs, the former were first submitted to the pre-processing steps as described in the previous section and then analysed with the respective MEM (see below). The MEMs used in each AA T version, together with a description of its model factors, are explained below:

AAT Arrow:

$$RBs = group + (1 \mid subject)\ (1)$$

This MEM was to analyse participants' general motor biases for pulling and pushing responses. In this model, participants RBs were modelled. The *group* factor refers to the two groups of participants, the LG or the HG. The term *(1|subject)* refers to the random effects.

AAT Training (RTs)**:**

$$RTs = group \times condition \times trial + (condition \mid subject/group) \quad (2)$$

This MEM was used to analyse participants RTs in the *approach negative* and *avoid neutral* conditions. In the model, *trial* refers to the number of times subjects performed the conditions throughout the five days, with 120 trials on each day; *condition* refers to the *approach negative* and *avoid neutral* conditions that participants trained; *group* refers to the two groups of participants, LG and HG. The term *(condition | subject/group)* refers to the random effects.

AAT Training (Ratings)**:**

$$Ratings = group \times condition \times session + (1 \mid subject/group) \quad (3)$$

This MEM was used to analyse the ratings of the negative and neutral-kitchen images that each participant trained with. In the model, *condition* refers to the two conditions (*approach negative* and *avoid neutral*) that all participants trained; *group* refers to the two groups of participants (LG and HG); *session* refers to the five training sessions after which participants had to rate the trained images. The term *(1 | subject/group)* refers to the random effects.

Reaction Biases (RBs) calculation:

The AAT assessment was implemented to capture the RB towards different stimuli at the 1st assessment and to perform comparisons to the RB displayed after training at the 2nd assessment. To analyse participants' RBs through their joystick reactions, the RTs obtained for the approach and avoid responses were used. This is one of the standard procedures that measures the strength of automatic associations towards specific stimuli, which involves using the subtraction between the mean RTs of all avoidance response for the stimuli of one category and the mean RTs of all approach responses for the stimuli of that same category (see formula 4). A positive value obtained indicates that the mean RTs for avoiding is slower (higher) than the mean RTs for approach, what is interpreted as an approach bias for the respective category. By the same reasoning, a negative value indicates an avoidance bias. However, an improved RB algorithm called '*D-score*' has been used in AAT studies[63,79,81,157] that standardizes the differences in the two mean RTs, by dividing these differences by the sum of standard deviations of the respective RTs (see formula 5). This procedure has been shown to reduce the influence of intra-individual variation in the RTs. Therefore, the current thesis used this improved algorithm to calculate the RBs for each participant. For example, for the MEM (1), the D-scores were obtained per participant for each category (negative, neutral-kitchen, neutral-street and positive) in each assessment session, using formula 5.

$$Bias = \overline{RTs_{avoid}} - \overline{RTs_{approach}} \quad (4)$$

$$D - score = \frac{\overline{RTs_{avoid}} - \overline{RTs_{approach}}}{\left(\frac{S_{avoid} + S_{approach}}{2}\right)} \quad (5)$$

AAT Assessment (RBs):

The following three MEMs were applied to address the hypotheses with regard to the training effects in implicit behaviour, in the AAT assessment:

$$Bias = group \times session \times category + (1 \mid subject/group) \quad (6)$$
$$Bias = group \times session \times trained + (1 \mid subject/group) \quad (7)$$
$$Bias = group \times session \times content + (1 \mid subject/group) \quad (8)$$

The three MEMs had in common the random effects term *(1 | subject/group)* and two factors, varying only in the third factor.

In all models, *group* refers to the two groups of participants, LG and HG; *session* refers to 1st and 2nd assessment session, before and after the training, respectively.

The third factor, depending on the MEM, allowed to capture differences in reaction biases between all four image categories (MEM 6 between trained and untrained images (MEM 7) or between different strengths of content in the negative images (MEM 8).

The MEM in (6) was used to analyse participants' RBs for each category per assessment session. In this model, *category* refers to the four distinct pictures categories, namely negative (dirty-toilets), positive (beach-scenarios), neutral-kitchen (kitchen environments) and neutral-street (street environments).

The MEM in (7) was used to analyse subject's RBs for the trained and untrained images, per assessment session. In more detail, this model was used to analyse subject's RBs for images in the negative and neutral-kitchen categories used in the training AAT version, separately. In this model, *trained* refers to the two negative and two neutral trained and to the two negative and two neutral untrained images pre-assigned to each participant. In particular, the former refers to the two AAT training images each participant saw during their training, while the latter refers to the other six AAT training images each participant only saw in the assessments.

The MEM in (8) was used to analyse subject's RBs for the assessment-only weak and strong images per assessment session. In this model *content* refers to the two types of content strengths of the negative images used in this model, namely the weak and strong, that were only shown in the AAT assessment versions.

With regard to the analysis procedure, the hypotheses and prerequisites were tested independently of higher significant interaction effects, the latter which were also analysed to not overlook any unexpected effects that could influence the interpretation of the results. P-levels were set to $p < 0.05$.

AAT Assessment (Ratings):

The following three MEMs were applied to address the hypotheses with regard to possible training effects on participants' explicit ratings, in the AAT assessment (These three models had an identical design to the ones used for the RBs in the AAT assessment, mentioned above):

$$Ratings = group \times session \times category + (1 \mid subject/group) \quad (9)$$
$$Ratings = group \times session \times trained + (1 \mid subject/group) \quad (10)$$
$$Ratings = group \times session \times content + (1 \mid subject/group) \quad (11)$$

Ratings Calculations

Besides analysing the RBs, the AAT assessment allowed to analyse how participants rated the same images they had to react to, in each assessment session. To better capture how each participant rated each set of stimuli relative to their overall average ratings, the Z-score method was used. This method guarantees the comparability of ratings between participants. For example, in the case of the MEM in formula 9, the Z-score in each category for each participant was calculated using formula 12, where the *Ratings* term is the average classification for all images of one category, the *Ratings Overall* correspond to the overall average classification for the images in all categories, and the term *S-Ratings Overall* is the standard deviation associated with the overall average.

$$Z - score = \frac{Ratings - \overline{Ratings_{Overall}}}{S_{Ratings_{Overall}}} \quad (12)$$

Practical Test (Sitting Test):

$$Time = group + (1 \mid subject) \quad (13)$$

The MEM in (13) was used to analyse the time it took participants to sit down and to press the space bar. In this model, the *group* factor refers to the two groups of participants, LG and HG. The term (*1 | subject*) refers to the random effects.

Practical Test (Ratings):

$$Ratings = group \times category + (1 \mid subject/group) \quad (14)$$

The MEM in (14) was used to analyse participants' ratings for the negative, neutral and positive images shown only in the practical test. In this model, the *group* factor refers to the two groups of participants, LG and HG. The term *(1 | subject/group)* refers to the random effects.

3.Results

Content:

- ✓ ***AAT Arrow***
- ✓ ***AAT Training***
- ✓ ***AAT Assessment***
- ✓ ***Assessment Ratings***
- ✓ ***Practical Test***

3.1 AAT Arrow

The hypothesis established for this AAT version was that the two groups would not have any general RB, i.e., neither an approach nor an avoidance bias.

Results of the MEM showed that the two groups did not differ in their RBs (Z = -1.22, p = 0.25).

To test the hypothesis mentioned above, the RBs were compared to a theoretical zero bias, in each group. In contrast with the expectations, results showed that the LG displayed an approach bias (Z = 2.21, p = 0.03). As for the HG no significant bias was found (Z = 0.76, p = 0.50).

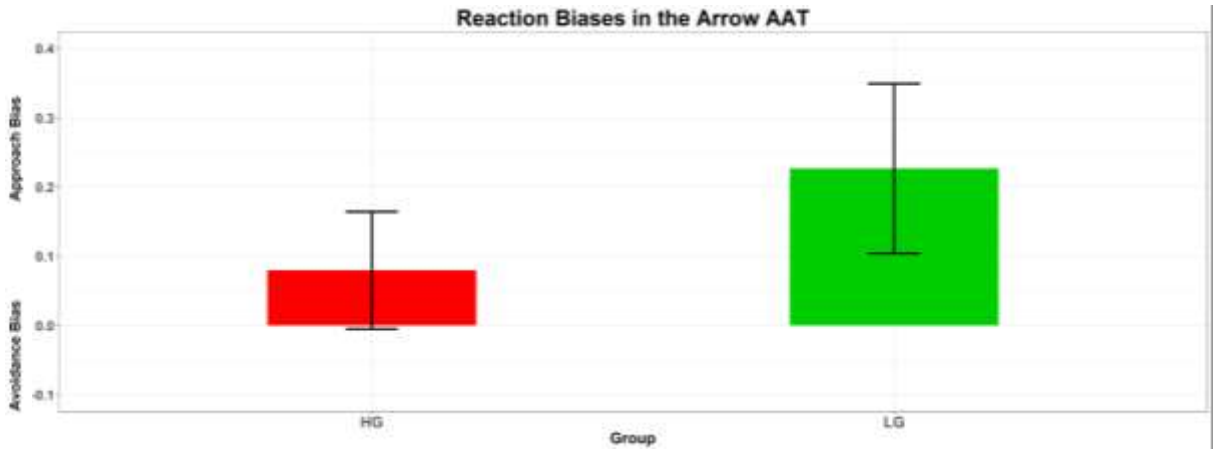


Figure 14: Reaction Biases in the AAT Arrow –Average reaction bias (RB; y-axis) that each group displayed when reacting to the direction of the arrows. An approach bias indicates that the reaction times for pulling the joystick is faster than pushing the joystick away. As a main result, when comparing the RBs of each group to a theoretical zero bias, the LG displayed an approach bias, while the HG did not display any bias. Importantly, the two groups did not differ from each other. Red – Group with high fear of contamination traits (HG). Green - Group with low fear of contamination traits (LG).

3.2 AAT Training

To understand how the RTs varied between conditions, groups and throughout the training period, MEMs were applied to estimate the influence of these three factors on the RTs, as described in more detail in the section *2.5 Mixed-Effects Models* in the Methods. For the up-coming section, we used the following order to describe the next set of results: intercept – interactions and main effects (condition, group); slope – interactions and main effects (condition*trial, group*trial).

3.2.1 Overview of Raw Reaction Times across Training

Prior to analysing the RTs obtained throughout the AAT training for the *approach negative* and *avoid neutral-kitchen* conditions, a brief overview of the performance of the conditions for each group was performed, in order to visualize the mean and variation of the RTs without any model fitting. All other analyses of the RTs throughout the training were performed upon fitting the RTs data with a power law.

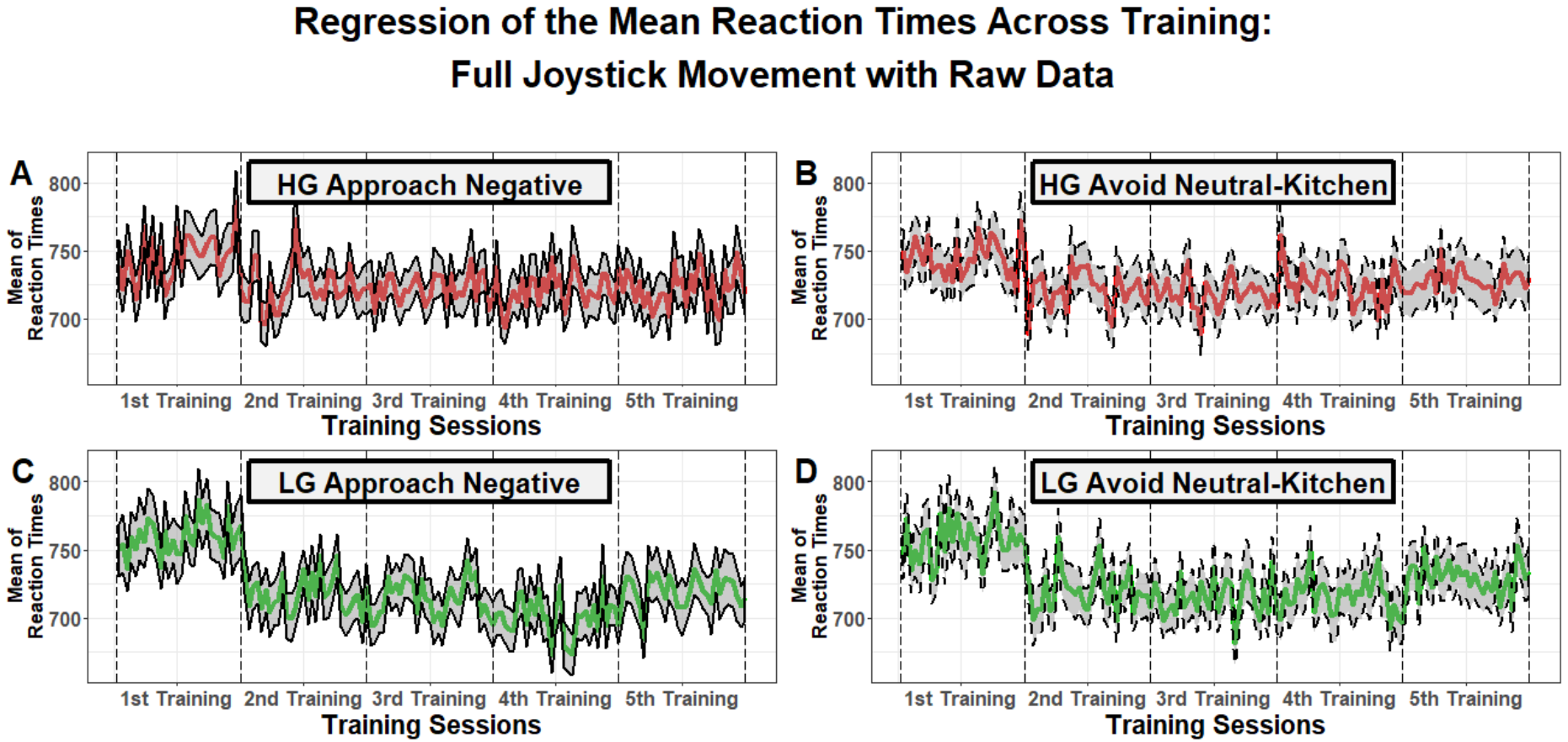


Figure 15: Full Joystick Movement for Raw Reaction Times – Each graph represents the average reaction times (RTs; y-axis) per trial that each group displayed for each condition, throughout the training period (x-axis). The two upper graphs (A and B) represent the RTs displayed by the group with high fear of contamination traits (HG; red lines), while the two lower graphs (C and D) represent the RTs displayed by the group with low fear of contamination traits (LG; green lines). The two left graphs (A and C) represent the approach negative condition (continuous black lines), while the two right graphs (B and D) represent the avoid neutral-kitchen conditions (dashed black lines). Black lines represent the 95% confidence interval.

3.2.2 Full Joystick Movement

With regard to the intercept of the training curves, i.e. the beginning of the training, the results of the MEM showed no significant differences (ME condition: F(1, 476.0) = 3.62, p = 0.06; ME group: F(1, 39.4) = 2.11, p = 0.15; IA condition x group: F(1, 725.7) = 3.31, p = 0.07). These findings led to the decision to not fix the intercepts, i.e., to not force the learning curves to start from the same point at trial = 1, to facilitate further analyses on the slopes. Such a procedure was not necessary here, since there were no systematic initial differences between groups and conditions.

Now focusing on the slopes of the training curves, for the hypothesis concerning the RTs between conditions in each group, it was expected that the HG would have a more pronounced speed-up of the RTs in the *approach negative* than in the *avoid neutral-kitchen* condition. The results of the MEM analyses for the two conditions revealed a main effect of trial (F(1.25058.1) = 304.4, p < 0.001): All conditions displayed a general downward slope, indicating that both groups got gradually faster throughout the training, in the two conditions. With regards to the 2-way interactions, results revealed an interaction between group and trial (IA group*trial F(1,25057.9) = 33.5, p < 0.001) (visual inspection of figure 17 revealed the character of the interaction to be ordinal): The LG displayed steeper slopes compared to the HG. In addition, there was an interaction between condition and trial (IA condition*trial: F(1.25056.1) = 6.52, p = 0.001): The condition *approach negative* displayed steeper slopes in both groups than *avoid neutral-kitchen*.

Lastly, MEM results revealed a significant 3-way interaction (IA group*condition*trial: F(1.25056.1) = 5.58, p = 0.02) (visual inspection of figure 17 revealed the character of the interaction to be ordinal, hence the main effect of trial – as reported above – was interpretable): the slopes of all four curves were decreasing at a different rate. To analyse this interaction in more detail, post-hoc tests were performed to compare the slopes between condition, in each group, and to compare the *approach negative* condition between groups. In the former contrasts, results showed that the LG speeded-up RTs in the *approach negative* condition more than in the *avoid neutral* condition (Z = 3.52, p < 0.001), whereas in the HG, no differences were found (Z = 0.13, p = 0.89). For the comparison of *approach negative* between groups, results showed that the LG speeded-up the RT in the *approach negative* condition more than the HG (Z = -5.74, p < 0.001). To check if both groups had speeded-up in the *approach negative* condition, additional contrasts tests were performed to compare the slopes of this condition, in each group, to a theoretical flat horizontal line. Results showed that both groups significantly speeded-up RTs in the *approach negative* condition (HG: Z = -5.88, p < 0.001; LG: Z = -14.06, p < 0.001).

To further analyse the RTs of the training, a MEM analysis was performed for two different subcomponents of the RTs: the time it took to initiate the joystick movement, which will now be referred to as *initiation RTs*, and the time it took to completely tilt the

joystick after movement had started, which will now be referred to as *motion RTs* (see figure 16). This splitting allowed to focus the analyses on the initial instants when participants had to decide which action to perform (*initiation RTs*), and on the rest of the joystick movement after the decision (*motion RTs*).

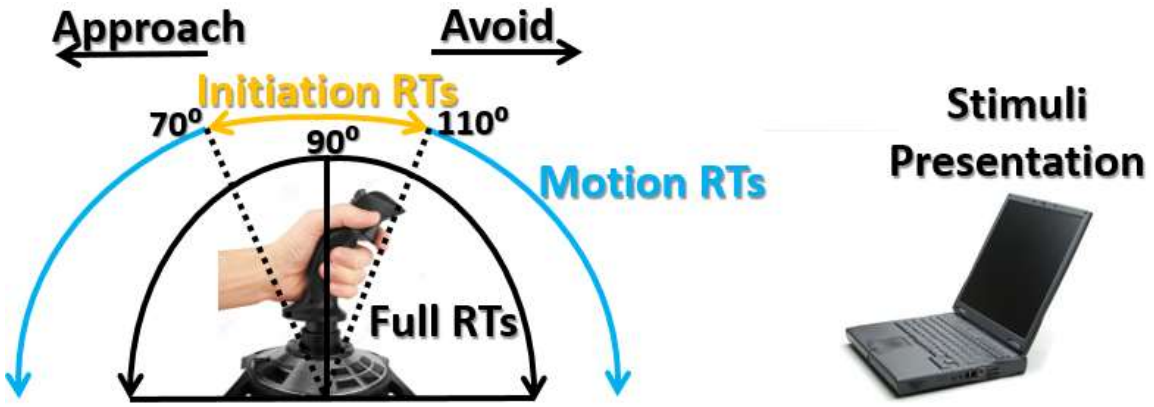


Figure 16: Reaction Times Subcomponents – The reaction times (RTs) of the full joystick movement (black arrows) were split into the initiation RTs and motions RTs. The initiation RTs consisted of the time it took for the joystick to reach the 20º degree angle when performing either an approach or an avoidance reaction, relative to the resting position at 90º degrees (yellow arrows), upon stimulus presentation. The motion RTs consisted of the time it took for the joystick to go from the 20º angle to the end of motion (blue arrows), when performing either an approach or an avoidance reaction. An approach response (joystick pull motion) would results in a zoom-in of the image on the screen, while an avoidance response (joystick push motion) would results in a zoom-out of the image on the screen.

3.2.3 Initiation RTs

With regard to the intercepts, the results of the MEM revealed a main effect of group ($F(1, 40.2) = 4.61$; $p = 0.04$) : The LG had initially higher RTs in both conditions, compared to the HG. In addition, MEM results showed a significant interaction between condition and group ($F(1, 534.7) = 5.90$; $p = 0.02$), for which post-hoc tests were performed to compare both conditions in each group. Results showed that the LG had initially higher RTs in the *approach negative* compared to the *avoid neutral-kitchen* condition ($Z = -2.83$, $p = 0.005$). As for the HG, no significant differences were found between conditions ($Z = 0.62$, $p = 0.54$).

Concerning the slopes of the training curves, MEM results revealed a main effect of trial ($F(1,25031.3) = 240.2$, $p < 0.001$): All slopes were generally decreasing, indicating that both groups got gradually faster in the two conditions. With regard to the 2-way interactions, MEM results showed an interaction between group and trial ($F(1, 25031.3) = 240.2$; $p < 0.001$), (visual inspection of figure 17 revealed the character of the interaction to be ordinal): The LG had steeper slopes compared to the HG.

Lastly, MEM showed that the 3-way interaction between group, condition and trial was significant (IA group*conditions*trial: $F(1, 25030.5) = 12.51$; $p < 0.001$) (visual

inspection of figure 17 revealed the character of the interaction to be ordinal, hence the main effect of trial – as reported above – was interpretable): the slopes of all four curves were decreasing at a different rate. To analyse this interaction in more detail, post-hoc tests were performed to compare the slopes between conditions within groups, and to compare the *approach negative* condition between groups. In the former contrasts, results showed that the LG speeded-up RTs in the *approach negative* condition more than in the avoid neutral condition ($Z = 3.70$, $p < 0.001$), while in the HG there were no differences ($Z = -1.33$, $p = 0.18$). In the contrasts used to compare the *approach negative* between groups, results showed that the LG speeded-up the RT in the *approach negative* condition more than the HG ($Z = -7.62$, $p < 0.001$). To check if both groups had speeded-up in the *approach negative* condition, additional contrasts tests were performed to compare the slopes of this condition, in each group, to a theoretical flat horizontal line. Results showed that both groups significantly speeded-up RTs in the *approach negative* condition throughout training (LG: $Z = -14.00$, $p < 0.001$; HG: $Z = -3.12$, $p = 0.002$).

3.2.4 Motion RTs

With regard to the intercepts, the results of the MEM revealed a significant 2-way interaction between conditions and groups ($F(1, 1324.6) = 7.10$; $p = 0.008$), but no systematic differences between groups ($F(1, 29.6) = 0.96$, $p = 0.34$). To explore the effects of the interaction between conditions and groups, post-hoc tests were performed to compare both conditions in each group. The results showed that the HG had initially higher motion RTs in the *approach negative* compared to the *avoid neutral* condition ($Z = -2.80$, $p < 0.001$). There were no significant differences in the LG ($Z = -0.96$, $p = 0.34$).

Concerning the slopes of the training curves, MEM results revealed a main effect of trial ($F(1, 22201.3) = 18.6$, $p < 0.001$): All slopes were generally decreasing, indicating that both groups got gradually faster in the two conditions. With regard to the 2-way interactions, MEM results showed an interaction between group and trial ($F(1, 25031.3) = 240.2$; $p < 0.001$) (visual inspection of figure 17 revealed the character of the interaction to be disordinal): The LG had a steeper slope only for the *approach negative* condition, compared to the HG. Moreover, MEM results showed an interaction between condition and trial ($F(1,22232.8) = 6.08$, $p = 0.014$): The *approach negative* condition had steeper slopes than *the avoid neutral-kitchen*.

Lastly, MEM showed that the 3-way interaction between group, condition and trial was significant (IA group*condition*trial: $F(1, 22232.8) = 13.8$; $p < 0.001$) (visual inspection of figure 17 revealed the character of the interaction to be disordinal, hence this main effect of trial – reported above - was not interpretable): the slopes of all four curves were not decreasing at a different rate. To analyse this interaction in more detail, post-hoc tests were performed to compare the slopes between conditions within

groups, and to compare the *approach negative* condition between groups. In the former contrasts, results showed that the HG speeded-up more in *approach negative* than in *avoid neutral* ($Z = 4.36$, $p < 0.001$), while in the LG, there were no differences between conditions ($Z = -0.88$, $p = 0.38$). In the contrasts used to compare the *approach negative* condition between groups, results showed that the HG speeded-up more than the LG ($Z = 5.294$, $p < 0.001$). To check if both groups had speeded-up in the *approach negative* condition, additional contrasts tests were performed to compare the slopes of this condition, in each group, to a theoretical flat horizontal line. Results showed that the HG significantly speeded-up in the *approach negative* condition ($Z = -7.10$, $p < 0.001$), as opposed to the LG ($Z = 0.40$, $p = 0.063$).

Thus, the findings for the whole joystick movement are in line with the previously established hypotheses, except for the one concerning the comparison between conditions in each group. In particular, the expected findings were the following: The RTs vary between conditions, groups and trials; both groups speeded-up in the *approach negative* condition throughout the training; the LG had steeper learning curves than the HG in the *approach negative* condition. In addition, the LG had speeded-up in the *approach negative* more than in *the avoid neutral* condition, as expected. Additionally, and more importantly, the main finding was that upon splitting the full joystick movement into two different subcomponents, the groups differed on these subcomponents throughout the training period. In particular, in the initiation movement, i.e., the time it took to start the joystick movement, the LG speeded-up in the *approach negative* condition more the HG. However, in the motion movement, i.e., the time it took to complete the joystick movement, the HG speeded-up more than the LG in the *approach negative* condition.

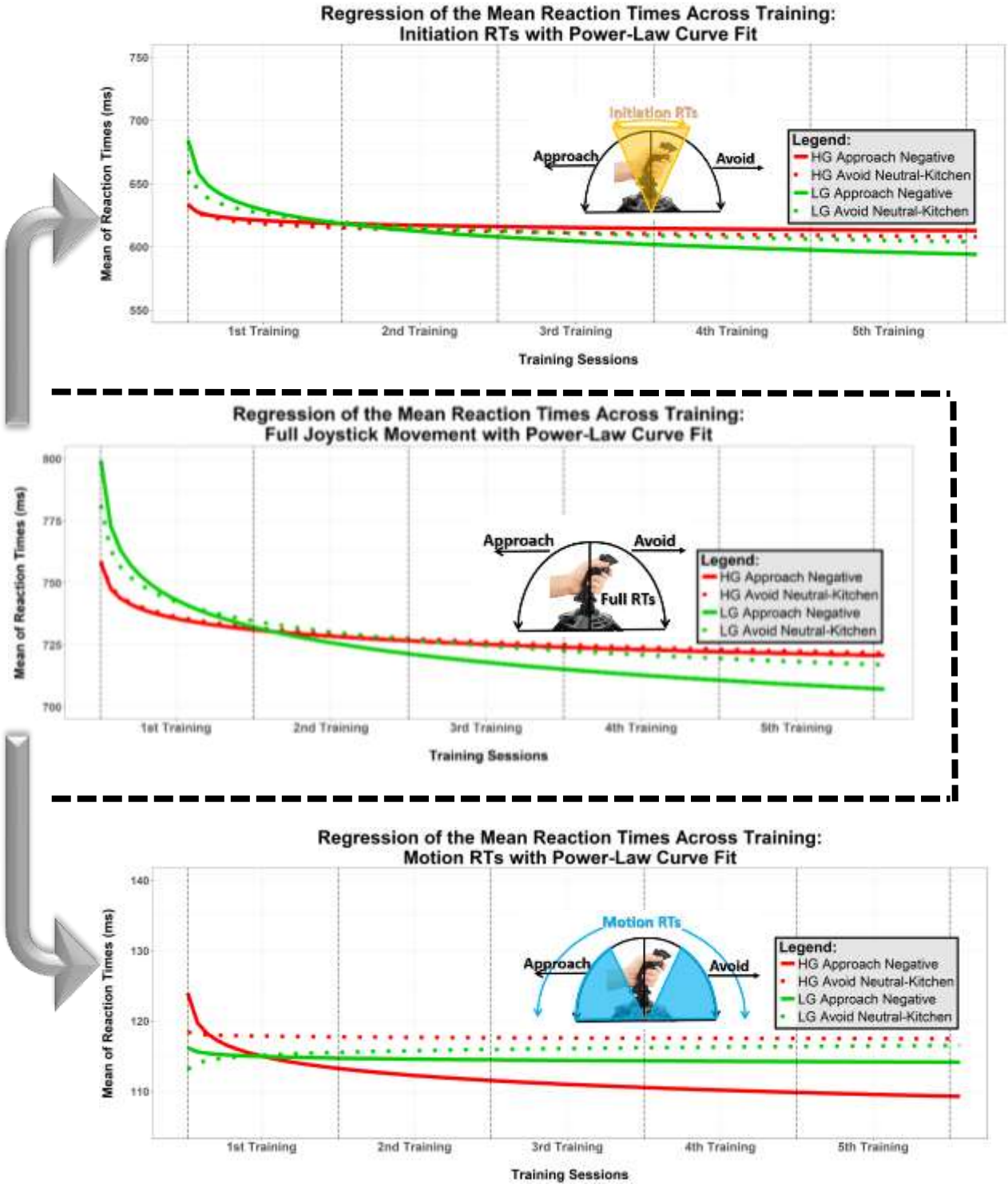


Figure 17: Reaction Times Performance in Training – Average reaction times (RTs; y-axis) fitted with a power law ($y = a \times x^b$) for the approach negative and avoid neutral condition displayed at each point throughout the training period (x-axis) by the group with high (HG; green lines) and the group with low (LG; red lines) fear of contamination traits. The full joystick movement (panel A) was decomposed in two parts, namely the time it took to initiate the joystick movement (initiation RTs, panel B) and the time it took to complete joystick movements (motion RTs; panel C). Main results were obtained by comparing approach negative between groups, for each of the three RTs. **Panel A** - The LG speeded-up in the full joystick movement more than the HG. **Panel B** - The LG speeded-up the initiation RTs in the approach negative condition more than the HG**. Panel C** – The HG speeded-up the motion RTs in the approach negative condition, more than the HG.

3.2.4 Ratings

The purpose of the ratings at the end of each training session was to have an explicit measure of participants' emotional bias towards the negative and neutral-kitchen trained pictures.

One important note was that since the ratings were obtained through the Z-score formula, which calculated participants' *average ratings* for the negative and neutral-kitchen images together, the average Z-score obtained in each participant for the neutral-kitchen images was significantly above zero (see figure below), as opposed to the raw rating values, exemplified in the *Overview of Raw Rating Scores* Table, in the section *3.4 Ratings* of the results.

Results of the MEM showed a main effect of condition ($F(1, 372.1) = 7286.6$, $p < 0.001$), while no other effects were reported (group: $F(1, 43.45) = 1.59$, $p = 0.21$, session: $F(4, 374.5) = 1.04$, $p = 0.38$; IA group*condition*session: $F(4, 372.1) > 0.99$, $p = 0.41$). Thus, these results indicate that the negative images were rated as significantly more unpleasant than the neutral-kitchen images, throughout all training sessions.

With regards to the hypothesis on a possible decrease in the unpleasantness ratings towards the negative images across the training days, contrast tests were performed to compare the negative images' ratings between the 1st and 5th training sessions, within groups. In contrast to the initial hypothesis, results revealed no rating differences between the 1st and the 5th training sessions, in none of the two groups, for the negative images (HG: $Z = 0.15$, $p = 0.88$; LG: $Z = -1.3$, $p = 0.19$).

Despite no significant interaction between session and group ($F(4, 374.54) = 0.52$, $p = 0.72$), an exploratory analysis was performed to compare the negative images' ratings between groups at the 5th training session. Contrast tests showed that the HG marginally rated the negative images as having elicited less unpleasant feelings, compared to the LG ($Z = 1.81$, $p = 0.07$).

Figure 18: Ratings Throughout the Training – Average ratings (y-axis) for the negative and neutral-kitchen images, performed at the end of each training session (x-axis). Each image was rated in terms of their "Unpleasant"/"Pleasant" characteristics, with a scale ranging from -100 (very unpleasant) to 100 (very pleasant). Dashed Lines – Images from the Avoid Neutral-Kitchen condition. Solid Lines – Images from the

Approach Negative condition. Red Lines – Ratings performed by the group with high fear of contamination traits (HG). Green Lines – Ratings performed by the group with low fear of contamination traits (LG).

Thus, findings of the ratings performed after each training session showed despite a marginal group differences on the ratings performed on the 5th training session, that throughout the training period both groups did not modify their unpleasant ratings towards the negative images.

3.3 AAT Assessment

In this AAT version, participants were presented images of different categories displaying either positive, neutral and negative content on a computer screen, for which participants had to perform approach and avoid reactions with a joystick, in an equally number of times. The purpose of the AAT Assessment was to have an implicit measure of participants' emotional bias towards all image categories (negative, neutral-kitchen, positive and neutral-street), before and after the training period. For a more detailed explanation of this version, see the *AAT versions* section in the Methods.

To understand how the RBs varied between different image categories, groups, assessment sessions, different content strengths and trained *versus* untrained images, different MEMs were applied to estimate the influence of these factors on the RBs, as described in more detail in the *Mixed-Effects Models* section in the Methods.

In addition, given the high level of noise in the data, due to inter and intra-individual RTs variability, trend effects are also reported.

3.3.1 Overview of Subjects' Performance before Training

Prior to examining the results obtained in the AAT assessment data through the respective MEMs, an overview was performed to examine how the criteria used to pre-select subjects, namely the obsessive-compulsive traits, influenced their performance before the 1sttraining session. To accomplish this, correlations between these traits with the RBs and the ratings were performed.

Due to the restricted diversity of subjects' washing subscale scores ranging from 1 (minimum) to 12 (maximum), clustering around 1 to 4 (see section *2.3 Sample Description*), the correlations were not performed with the Washing Subscale scores, but instead with the OCI-R scores.

In addition, these analyses were executed specifically for the images in the negative category, since this was the main focus of the hypotheses established for the AAT assessment.

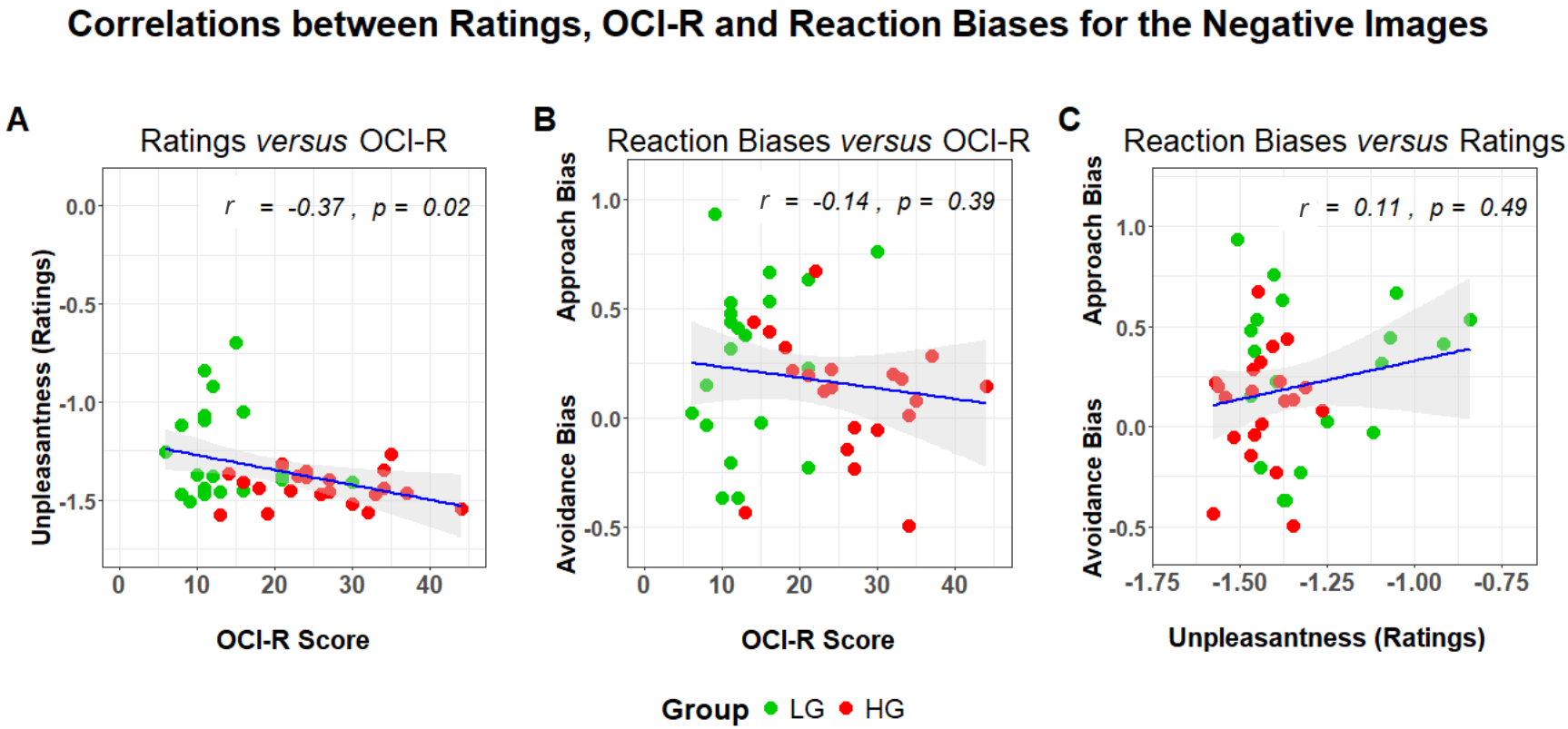


Figure 19: Correlation between Ratings, OCI-R and Reaction Biases for Negative Images, before Training. **Left Panel** - Average ratings (y-axis) for all negative images by each participant (dots) at the 1st assessment, with their respective OCI-R scores (x-axis). This correlation indicated that the higher the OCI-R scores, the more unpleasant participants rated the negative images (r = -0.39, p = 0.01). **Middle Panel** - Average reaction biases (RBs; y-axis) for all negative images by each participant (dots) at the 1st assessment, with their respective OCI-R scores (x-axis). This correlation was not significant (r = -0.14, p = 0.39), indicating that the individual differences in the OCI-R scores did not explain the RBs displayed for the negative images. **Right Panel** - Average RBs (y-axis) by each participant (dots) at the 1st assessment, with their respective ratings (x-axis), for the negative images. This correlation was not significant (r = 0.11, p = 0.49), indicating that the individual differences in the OCI-R scores did not have an influence of the RBs for the negative images. Red dots – participants from the high fear of contamination-related trait group (HG). Green dots - participants from the low fear of contamination-related trait group (LG).

The hypotheses for these brief analyses were that the higher the OCI-R score was, the more unpleasant subjects would rate and the stronger their avoidance RB would be towards the negative images. For the correlation between the negative images' ratings and the OCI-R scores, the results showed, as expected, a negative correlation between the ratings of the negative images and the OCI-R score of both groups ($r = -0.39$, $p = 0.02$), indicating that the higher subjects scored in the OCI-R scale, the more unpleasant they rated the negative images. On the contrary, no significant correlations were found between the RBs in the negative images with the OCI-R scores ($r = -0.14$, $p = 0.39$) nor between the RBs with the ratings ($r = 0.11$, $p = 0.49$) of the negative images. Respectively, this indicated that participants' OCI-R scores did not explain the individual RBs differences displayed for the negative images, and that these RBs differences were not associated with systematic differences in the unpleasantness towards the negative images.

3.3.2 Comparisons of Reaction Biases Between Image Categories

Results of the MEM analysis for all image categories showed no significance in the 3-way interaction (IA group*session*category $F(3, 301) = 0.34$, $p = 0.80$) nor in the effects (category: $F(3, 301) = 2.10$, $p = 0.10$; session: $F(1, 301) = 0.58$, $p = 0.45$; group: $F(1, 43) = 2.29$, $p = 0.14$)).

Before testing the hypotheses initially established for these analyses, both groups' initial RBs were analysed in each category at the 1st assessment, since this allowed to examine if the prerequisites were met, and, consequently, to correctly interpret the results afterwards. To test these prerequisites, contrast tests were performed to compare the RBs elicited by all images in each category to a theoretically null bias, on the 1st assessment, in each group. The aim here was to check whether the images had – before any training –elicited RBs as expected in both groups: Negative images should have elicited an avoidance bias, positive images an approach bias and the neutral images should not have elicited any specific RB. This analysis showed that the LG displayed a significant approach RBs for the images in the negative category ($Z = 2.04$, $p = 0.041$) in contrast to the expectations, no significant RB for the images in the neutral-kitchen (Neutral-Kitchen: $Z = 1.18$, $p = 0.24$), and a marginal approach RB for the images in the neutral-street positive categories (Neutral-Street: $Z = 1.79$, $p = 0.073$; Positive: $Z = 1.72$, $p = 0.086$). The HG did not display significant RBs for any category (negative $Z = 1.18$, $p = 0.24$; neutral-kitchen $Z = -0.24$, $p = 0.81$; neutral-street $Z = -0.49$, $p = 0.62$ and positive $Z = -0.02$, $p = 0.98$). In light of these results, particularly with regard to the approach RBs in the LG for images in the negative category, the pre-requisites were only partly fulfilled.

Additionally, as an extension of the prerequisites tested above, contrast tests were performed to compare the RBs elicited by the negative images *between* groups, at

the 1st assessment. Due to their higher sensitivity for the negative stimuli, it was expected that the HG would show a stronger avoidance RB than the LG. Taking into account the analysis above, these results showed, in contrasted to what was expected, that the groups did not differ in their RBs for the images in the negative category, on the 1st assessment (Z = -0.58 p = 0.56). In addition, when applying the same between-group comparison to the other three categories, results indicated that the groups also did not differ in the RBs for images in the neutral-kitchen (Z = -1, p = 0.32), neutral-street (Z = -1.61, p = 0.11) and positive categories (Z = -1.21, p = 0.23), at the 1st assessment.

Regarding the hypothesis about the RBs elicited by the images in the negative category from the 1st to the 2nd assessment, between groups, it was predicted that the LG would have a stronger avoidance RB reduction between assessments than the HG, due to the LG's lower sensitivity for negative stimuli. No such interaction [(2nd Assess. – 1st Assess., in HG) – (2nd Assess. – 1st Assess., in LG)] between session and group was observed for the negative images (Z = -0.65, p = 0.52; see figure 20). Additionally, when applying the same interaction contrast for the other three categories, results showed that groups also did not differ in the alterations of the RBs for the images in neutral-kitchen (Z =-0.38, p = 0.70), neutral-street (Z = -0.20, p = 0.85) and positive categories (Z = 0.69, p = 0.49).

To not miss weaker alterations of RBs, RB alterations were then tested *within* groups [2nd Assess. – 1st Assess.], separately for all image categories. This allowed to examine if each group had actually displayed RB differences across the assessments as initially thought, particularly for the negative images. Here the expectation was that both groups would have decreased their avoidance RB for the negative images, increased their approach RB for the neutral-kitchen images and not have any RB alteration for the neutral-street and positive images. Results showed that LG did not display RB differences between assessments in any category (negative: Z = 0.60, p = 0.55; neutral-street: Z = 0.96, p = 0.34; neutral-kitchen: Z = -0.19, p = 0.85; positive: Z = -0.44, p = 0.66). Similarly, the HG did not display any RB alterations (negative: Z = -0.33, p = 0.75; neutral-street: Z = 0.67, p = 0.51; neutral-kitchen: Z = -0.19, p = 0.85; positive: Z = 0.53, p = 0.60).

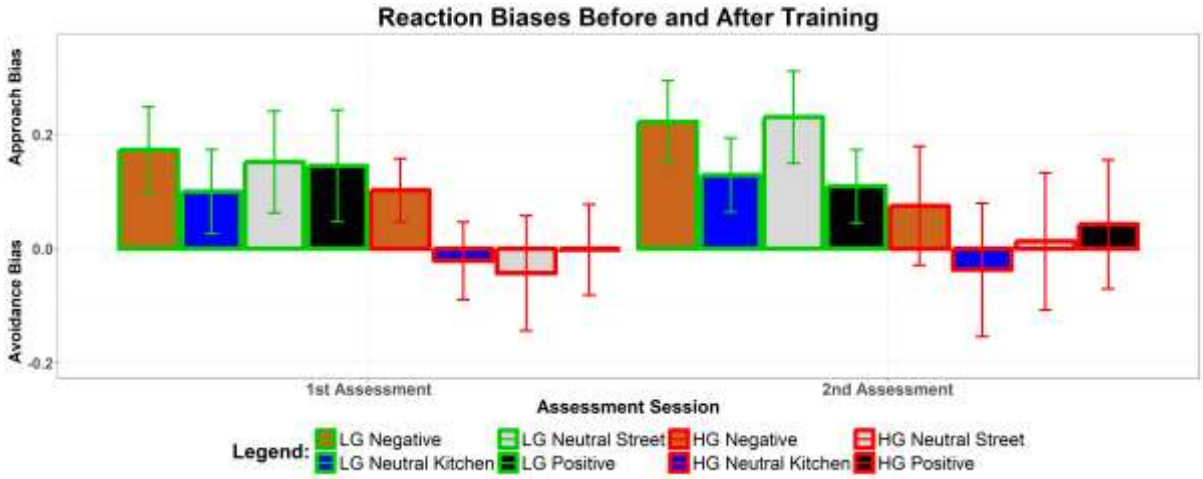


Figure 20: Reaction Biases before and after Training – Average Reaction Bias (RBs; y-axis) for each category of images, i.e., negative (brown), neutral-kitchen (blue), positive (black) and neutral-street (light grey), displayed by the group with low (green outline; LG) and the group with high (red outline; HG) fear of contamination traits at the 1st and 2nd assessments (x-axis). Main results were obtained by comparing the RB variation across the assessments displayed for the negative category between groups, which showed that the two groups did not differ in this comparison.

3.3.3 Comparison of Reaction Biases for Trained vs Untrained Images

Results of the MEM analysis for the medium content negative images used in the AAT training showed no significance in the 3-way interaction and no main effects in the RBs (IA group*session*trained: $F(1, 129) = 0.18$, $p = 0.67$; session: $F(1, 129) = 1.09$, $p = 0.30$; group: $F(1, 43) = 1.42$, $p = 0.24$; trained: $F(1, 129) = 0.03$, $p = 0.86$). With regard to the 2-way interactions, results showed a trend in the interaction between session and trained/untrained negative images (IA session*trained: $F(1,129) = 2.98$, $p = 0.09$), for which visual inspection of the figure below suggested that the effect was driven by a more pronounced increase in the approach RB from the 1st to the 2nd assessment for the negative trained images, compared to the untrained ones. In addition, MEM results revealed an interaction between session and group (IA session*group: $F(1,129) = 4.85$, $p = 0.03$). To analyse this interaction in more detail, post-hoc tests were performed to compare both groups separately at the 1st and 2nd assessment, and to compare the RBs between assessments in each group. The between-group tests showed no group difference in RBs for the negative images at the 1st assessment ($Z = 0.40$, $p = 0.69$), but showed that the LG displayed stronger approach RBs for the negative images than to the HG, at the 2nd assessment ($Z = -2.29$, $p = 0.02$). The within-group tests showed that the LG displayed a stronger approach bias at the 2nd assessment compared to the 1st assessment ($Z = 2.32$, $p = 0.02$), where no differences were found in the HG ($Z = -0.81$, $p = 0.42$). As such, these four contrast tests indicated that the interaction between session and group was driven mostly by an increase in the approach RB from the 1st to the 2nd assessment in the LG, whereby there were no changes in the HG.

In addition, contrast tests were performed to compare the RBs of the negative images for trained *versus* untrained *at the 1st assessment*, in each group, i.e., when no training had been performed yet. More specifically, the aim here was to check if both groups had displayed similar RBs at the 1st assessment when comparing the trained to the untrained images, since both had an *equal* medium content strength (See figure 8 for a better understanding, in section *2.2.1.3 AAT Training* in the Methods). As expected, results showed that none of the groups displayed RB differences in the trained *versus* untrained negative images (LG: $Z = -0.63$, $p = 0.53$; HG: $Z = -1.26$, $p = 0.21$).

Regarding the neutral-kitchen images, results of the MEM analysis for the neutral-kitchen images used in the AAT training showed no significance in the 3-way interaction (IA: session*group*trained: $F(1,129) = 0.09$, $p = 0.77$) and no main effects (session: $F(1,129) = 0.11$, $p = 0.74$; group: $F(1,43) = 2.69$, $p = 0.11$; trained: $F(1,129) =$

0.62, p = 0.43). With regard to the 2-way interactions, MEM results showed a trend in the interaction between session and trained/untrained neutral-kitchen images (IA session*trained F(1,129) = 2.98, p = 0.09), for which visual inspection of the plot suggests that the effect was driven by an approach RB increase from the 1st to the 2nd assessment more for the trained, than the untrained neutral-kitchen images.

In addition, contrast tests were performed to examine the RB for neutral-kitchen trained *versus* untrained images *at the 1st assessment*, in each group, in a similar fashion as described above for the negative trained versus untrained images. As expected, results showed that none of the groups displayed RB differences in the trained *versus* untrained neutral-kitchen images (LG: Z = -1.24; p = 0.22; HG: Z = -1.26, p = 0.21).

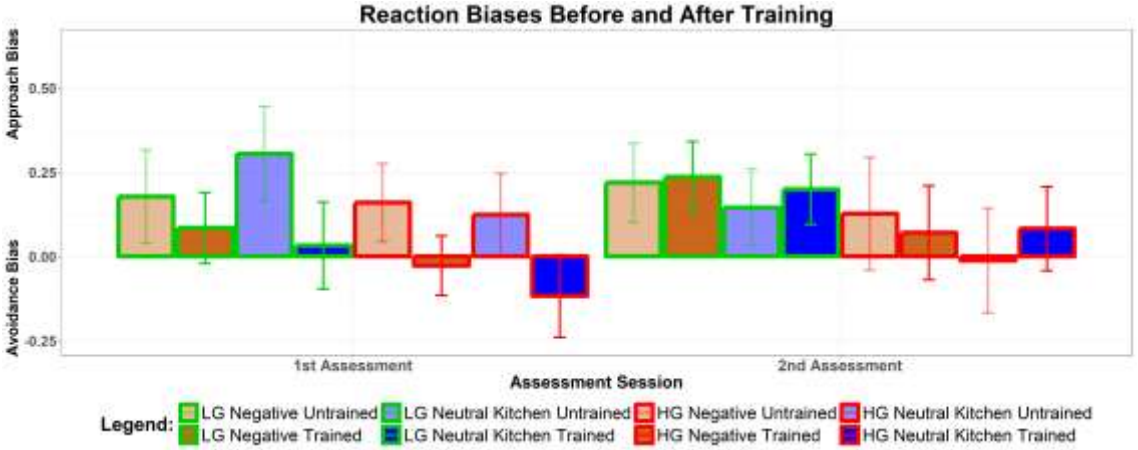


Figure 21: Reaction Biases before and after Training – Average Reaction Bias (RBs; y-axis) for the negative (brown) and neutral-kitchen (blue) images, pseudo-randomly pre-assigned for each individual to be untrained (lighter colour) or trained (darker colour), at the 1st and 2nd assessment (x-axis). The RBs are shown for the two groups, i.e., for the group with low (green outline; LG) and the group with high (red outline; HG) fear of contamination traits.

3.3.4 Comparison of Reaction Biases for the Weak vs Strong Negative Images

Results of the MEM analysis for the negative weak *versus* strong images showed no significance in the 3-way interaction (IA *group*session*content*: F(2, 215) = 0.29, p = 0.75). MEM results revealed a trend effect in content strength (F(2, 215) = 2.54, p = 0.08), for which visual inspection of the figure below suggested that the effect was due to a stronger approach RB for the negative strong images compared to the weak negative images in both groups. In addition, regarding the 2-way interactions, results showed a trend for an interaction between session and group (IA session*group: F(1, 215) = 3.12, p = 0.08), for which visual inspection of the figure below suggested that the effect was due to a stronger increase of the approach RB from the 1st to the 2nd assessment in the LG, than in the HG. No other significant main effects were reported (session: F(1, 215) = 1.26, p = 0.264; group: F(1, 43) = 2.28, p = 0.144).

In addition, the 2-way interaction between content and group at the 1st assessment was analysed, since, originally, the expectation was to find a stronger avoidance bias for the strong than for the weak images, whereby, this pattern should be

more pronounced in the HG than in the LG. However, this interaction showed no effect (Z = 1.07, p = 0.29).

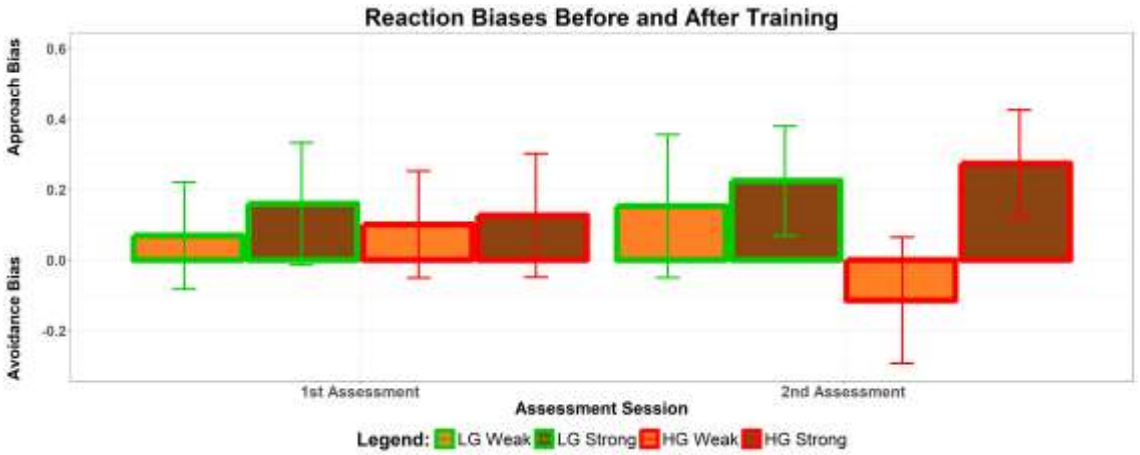


Figure 22: Reaction Biases before and after Training – Average Reaction Bias (RBs; y-axis) for the negative weak (lighter brown) and negative strong (darker brown) images, displayed by the group with low (green outline; LG) and the group with high (red outline; HG) fear of contamination traits at the 1st and 2nd assessments (x-axis). Main results showed no RBs difference in the negative weak versus strong images.

3.3.5 Comparison of Reaction Biases for Generalization Assessment

In order to measure the training generalization effects relative to the assessment-only negative images, i.e., compared to the negative weak and strong images, a correlation was performed with the RB differences, between the two assessments, in the negative weak and strong *versus* the negative medium trained images with the OCI-R scores. Upon inspection of each participant's RB changes, the value obtained for participant #42 was noted to be extremely high relative to the rest of the participants. As such, this participant was excluded from this analysis. Results (see figure 23) showed that the OCI-R score did not explain the individual RB changes in the medium trained negative *versus* weak negative images (r = -0.03, p = 0.87), neither in the strong *versus* medium trained negative images (r = -0.08, p = 0.61).

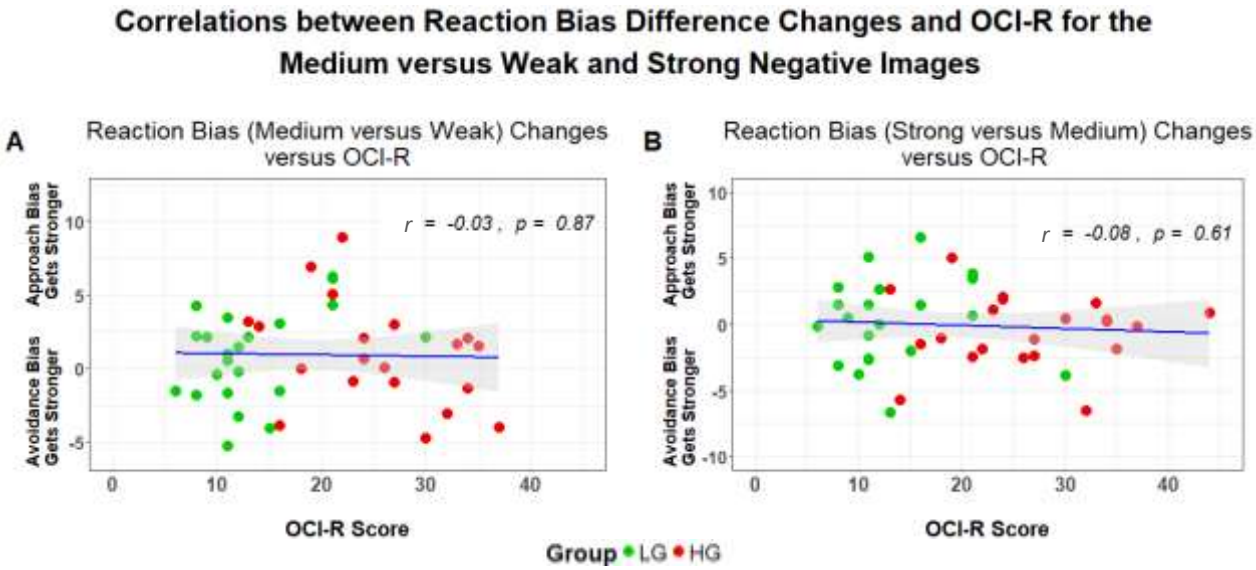


Figure 23: Correlation between Reaction Bias Difference Changes and the OCI-R for the Negative Images. **Left Panel** - Average reaction biases (RBs; y-axis) changes (between assessments) for the medium trained versus weak negative images differences, with the OCI-R scores (x-axis). **Right Panel** - Average RBs (y-axis) changes for the strong versus medium trained negative images differences, with the OCI-R scores (x-axis). Each dot corresponds to the contrast value obtained for each subject and the respective OCI-R score.

3.4 Ratings

In these ratings, participants were presented pictures of different categories separately displaying positive, neutral and negative characteristics on a computer screen, for which participants had to rate them on an "*Unpleasant*"/"*Pleasant*" scale. The purpose of the ratings realized before and after the joystick part, in the 1st and 2nd assessment sessions, respectively, was to have an explicit measure of participants' emotional bias towards all image categories (negative, neutral-kitchen, positive and neutral-street) before any training and after the five-day consecutive training period. For a more detailed explanation of these version, see the *AAT Versions* section, in the Methods.

To understand how the ratings varied between different image categories, groups, assessment sessions, different content strengths and trained *versus* untrained images, different MEMs were applied to estimate the influence of these factors on the ratings, as described in more detail in the section *2.5 Mixed-Effects Models* in the Methods.

3.4.1 Overview of Raw Ratings before Training

Prior to examining the results obtained in the ratings before and after the joystick part, at the 1st and 2nd assessment, respectively, an overview of the raw rating scores was performed to get a first impression of the range of the ratings (minimum score possible: -100, maximum score possible: +100). For the negative images, the raw ratings were obtained for the assessment-only images (weak and strong) and for the medium content images. As observable in the figure below, participants used a wide range of the possible rating scores, whereby, neither the positive nor the negative pictures caused ceiling effects.

Overview of Raw Rating Scores in each Category

		LG		HG	
Category		$\bar{X}$	$S.E.$	$\bar{X}$	$S.E.$
Positive		73.10	3.43	70.35	5.36
Neutral-Street		6.40	1.77	12.16	3.99
Neutral-Kitchen		16.54	4.00	34.20	5.61
Negative	Weak	-57.41	4.47	-68.86	2.43
	Medium	-64.73	3.49	-77.45	4.28
	Strong	-59.57	6.63	-84.7	3.20

Figure 24: Overview of Raw Ratings Score – Average ratings and respective standard error of the mean values for each category of images, before the joystick part at the 1st assessment. The column in green represents the ratings performed by the low fear of contamination traits' group (LG), while the column in red represents the ratings performed by the low fear of contamination traits' group (HG). In the negative category, the values were obtained for the negative weak, strong and medium content images.

For all following statistical analyses, the rating scores were obtained through the Z-score formula and analysed with the respective MEMs, as explained in more detail in the Methods.

3.4.2 Comparison of Ratings Between Image Categories

Results of the MEM analysis showed a main effect of category ($F(3,344) = 1678.4$, $p = < 0.001$), for which visual inspection of figure 25 evidences a clear difference in how each category was rated on average. With regards to the 2-way interaction, MEM results revealed an interaction between category and group (IA category*group: $F(3,344) = 20.35$, $p < 0.001$), for which visual inspection of the figure below suggested that the effect was due to the LG, in general, rating the images as being more pleasant, than the HG. No other significant effects were reported (IA group*session*category $F(3,344) = 0.82$, $p = 0.48$; group: $F(1,344) < 0.001$, $p > 0.99$; session: $F(1,344) < 0.01$, $p > 0.99$).

Before testing the hypotheses initially established for these analyses, both groups' initial ratings were analysed in each category at the 1st assessment, since this allowed to examine if the prerequisites were met, and, consequently, to correctly interpret the results afterwards. To test these prerequisites, contrast tests were performed to compare the ratings towards all images in each category to a theoretical zero (neither unpleasant nor pleasant), on the 1st assessment, in each group. The aim here was to check whether each group had rated - before any training - the images as expected in both groups: Negative images should have been rated as unpleasant, positive images as pleasant and the neutral images as neither pleasant nor unpleasant. This analysis showed, as expected, that both groups rated the negative images as unpleasant and the positive as pleasant (HG negative: $Z = -29.1$, $p < 0.001$; HG positive: $Z = 19.9$, $p < 0.001$; LG negative: $Z = -26.2$, $p <.001$; LG positive: $Z = 24.8$, $p < 0.001$). As for the neutral images, whereas both groups rated the neutral-street images as neither pleasant nor unpleasant (HG: $Z = 1.17$, $p = 0.24$; LG: $Z = -0.92$, $p = 0.36$), the neutral-kitchen images, however, were rated by both groups as being pleasant (HG: $Z = 8.08$, $p < 0.001$; LG: $Z = 2.40$, $p = 0.017$). Thus, the prerequisites tested were closely fulfilled.

Additionally, as an extension of the prerequisites tested above, contrast tests were performed to compare the ratings of the negative images *between* groups, at the 1st assessment. Due to their higher sensitivity for the negative stimuli, it was expected the HG would rate the negative images as being more unpleasant, than the LG. In line with the expectations, and taking into account the analysis above, the results showed that the HG rated the negative images as being more unpleasant, than the LG ($Z = -2.51$, $p = 0.012$). In addition, when applying the same between-group comparison to the other

three categories, results indicated that for the positive images the HG rated them as being less pleasant than the LG (Z = -3.08, p = 0.002). As for the neutral images, while there were no group differences for neutral-street images (Z = 1.48, p = 0.14), for the neutral-kitchen images, the HG rated them as more pleasant than the LG (Z = 4.1, p < 0.001).

Regarding the hypothesis about the rating of the images in the negative category from the 1st to the 2nd assessment, *between* groups, it was predicted that the LG would have a stronger unpleasantness reduction *across* assessments than the HG, due to the LG's lower sensitivity for negative stimuli. Contrast tests [(2nd Assess. – 1st Assess., in HG) – (2nd Assess. – 1st Assess., in LG)] showed. that the two groups did not differ in the variation of the ratings for the images in the negative category, across assessments (Z = -0.63, p = 0.53). Additionally, when applying the same contrast tests for the other three categories, results showed that groups also did not differ in the rating alterations for the images in the neutral-kitchen (Z = -0.92, p = 0.36), neutral-street (Z = 0.68, p = 0.50) and positive (Z = 0.87, p = 0.38) categories.

To not miss weaker alterations of ratings, rating alterations were then tested *within* groups [2nd Assess. – 1st Assess.], separately for all image categories. This allowed to examine if each group had actually displayed rating differences across the assessments as initially thought, particularly for the negative images. Here the expectation was that both groups would have decreased their unpleasantness for the negative images, increased their pleasantness for the neutral-kitchen images and not have any rating alterations for the neutral-street and positive images since these two were never trained. For both groups, while there were no rating alterations for the images in the neutral-street (HG: Z = -0.39, p = 0.7; LG: Z = -1.37, p = 0.17) and positive (HG: Z = 1.09, p = 0.27; LG: Z = -0.13, p = 0.9) categories as expected, there were no rating alterations in the negative (HG: Z = 0.66, p = 0.51; LG: Z = 1.58, p = 0.11) and neutral-kitchen (HG: Z = -1.37, p = 0.17; LG: Z = -0.085, p = 0.93) categories, in contrast to the expectations.

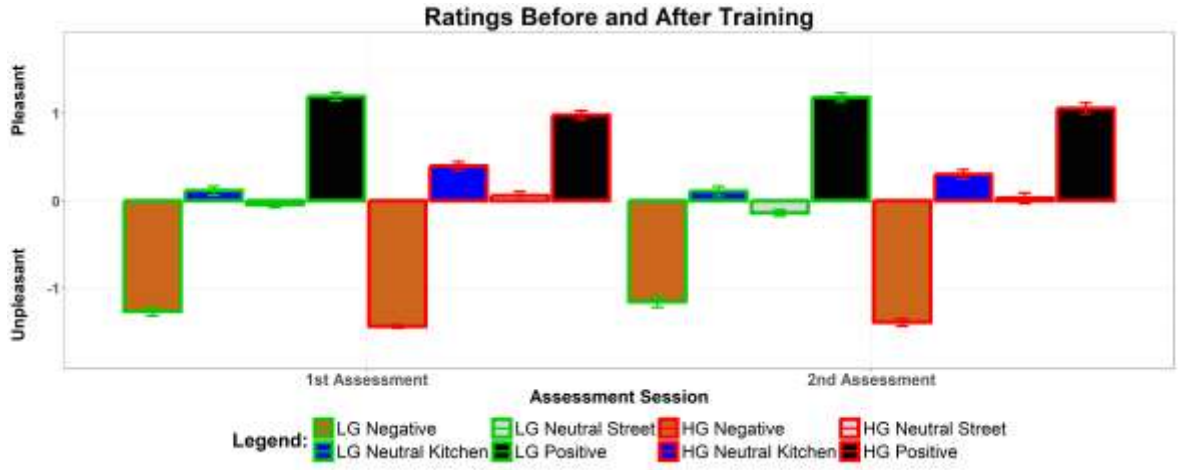


Figure 25: Ratings before and after Training – Average ratings (y-axis) for each category of images, i.e., negative (brown), neutral-kitchen (blue), positive (black) and neutral-street (light grey), for the group with low (green outline; LG) and the group with high (red outline; HG) fear of contamination traits at the 1st

and 2nd assessments (x-axis). Main results were obtained by comparing the ratings alteration across the assessments displayed for the negative category between groups, which showed that the two groups did not differ in this comparison. In addition, upon comparing the ratings between groups at the 1st assessment, results showed that the HG rated the negative category as being more unpleasant than the LG.

3.4.3 Comparison of Ratings for Trained versus Untrained Images

Results of the MEM analysis of the images used in the AAT training showed a main effect of session and a main effect of trained vs untrained for the negative images (session: $F(1,129) = 22.27$, $p < 0.001$; trained: $F(1,129) = 7.06$, $p = 0.009$), which were not interpretable upon visual inspection of the figure 27. A main effect of group was reported (group: $F(1,43) = 5.12$, $p = 0.003$), for which visual inspection of figure 27 revealed that the effect was driven by the fact that the HG rated the negative trained and untrained images as being more unpleasant in both assessments, compared to the LG. With regard to the 2-way interactions, MEM results showed an interaction between session and trained/untrained (IA session*trained $F(1,129) = 6,80$, $p = 0.01$). To analyse this interaction in more detail, post-hoc contrast tests were performed to compare the negative trained *versus* untrained images' ratings separately at the 1st and 2nd assessment, and to compare the negative trained *versus* untrained rating changes, between the 1st to the 2nd assessment. When taking both groups together, the within-assessment contrast tests showed that, at the 1st assessment, there were no trained *versus* untrained differences in the negative ratings ($Z = 0.18$, $p = 0.86$), while at the 2nd assessment the negative trained images were rated was being less unpleasant than the negative untrained images ($Z = 3.87$, $p < 0.001$). The between-assessment analyses showed that the negative trained images were rated as being less unpleasant at the 2nd assessment compared to the 1st assessment ($Z = 5.36$, $p < 0.001$), while the negative untrained images were marginally rated as also being less unpleasant at the 2nd assessment ($Z = 1.67$, $p = 0.09$). As such, these contrast tests showed that the interaction between session and trained was driven mostly by changes in the negative trained images from the 1st to the 2nd assessment, more than by the untrained images.

In addition, contrast tests were performed to examine the negative images' trained *versus* untrained rating differences *at the 1st assessment*, in each group, i.e., when no training had been performed yet. More specifically, the aim here was to check if both groups had displayed similar ratings at the 1st assessment when comparing the trained to the untrained images, since both had an *equal* medium content strength (See figure 26 for a better understanding, in the Methods). As expected, results showed that none of the groups displayed trained *versus* untrained rating differences in the negative images (HG: $Z = 0.47$, $p = 0.64$; LG: $Z = -0.21$, $p = 0.83$).

In an exploratory analysis, subjects' ratings for the negative trained images across the two assessments [2nd Assess. - 1st Assess] were correlated with their respective OCI-R scores. This analysis revealed a non-significant negative correlation (r = -0.2, p = 0.19), indicating that inter-individual differences with regard to the OCI-R score did not explain any rating changes in the negative trained images from the 1st to the 2nd assessment.

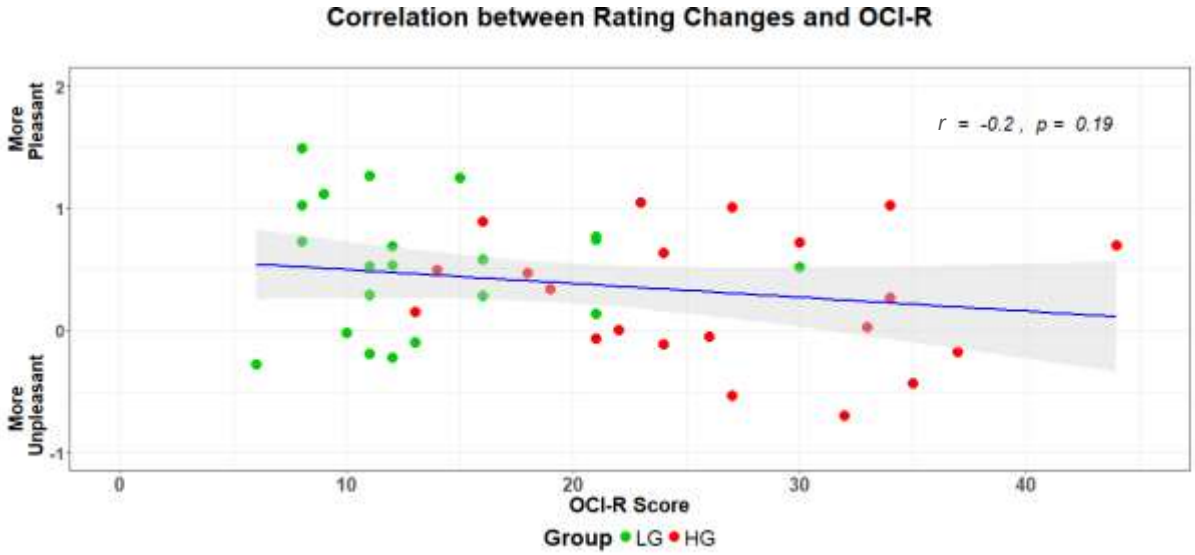


Figure 26: Correlation Between Ratings Changes and OCI-R - Average ratings (y-axis) changes between assessments for all negative images by each participant (dots), with their respective OCI-R scores (x-axis). This correlation revealed that the OCI-R did not explain any rating changes for the negative trained images between the 1st and the 2nd assessment (r = -0.2, p = 0.19).

Regarding the neutral-kitchen images, results of the MEM analysis for the neutral-kitchen images used in the AAT training showed a main effect of session (F (1,129) = 4.12, p = 0.04), which was not interpretable upon visual inspection of the figure below. In addition, MEM results revealed a main effect of group (F(1,43) = 13.04, p < 0.001), for which visual inspection of the figure indicated that the effect was driven by the fact that the neutral-kitchen images were always rated as being more pleasant by the HG, compared to the LG. With regard to the 2-way interactions, a trend was reported for the interaction between session and group (IA session*group: F(1,129) = 2.90, p = 0.09), for which visual inspection of the figure below suggested that the marginal effect was driven by a decrease in pleasantness in the HG from the 1st to the 2nd assessment, compared to the LG. No other significant effects were reported (trained: F(1,129) = 1.51, p = 0.22; IA session*group*trained: F(1,129) = 0.34, p = 0.56).

In addition, contrast tests were performed to examine the neutral-kitchen images' trained *versus* untrained rating differences *at the 1st assessment*, in each group, in a similar fashion as described above for the negative trained *versus* untrained images' RBs. As expected, results showed that none of the groups displayed rating differences in the trained *versus* untrained neutral-kitchen images (LG: Z = -1.02, p = 0.31; HG: Z = -0.56, p = 0.57).

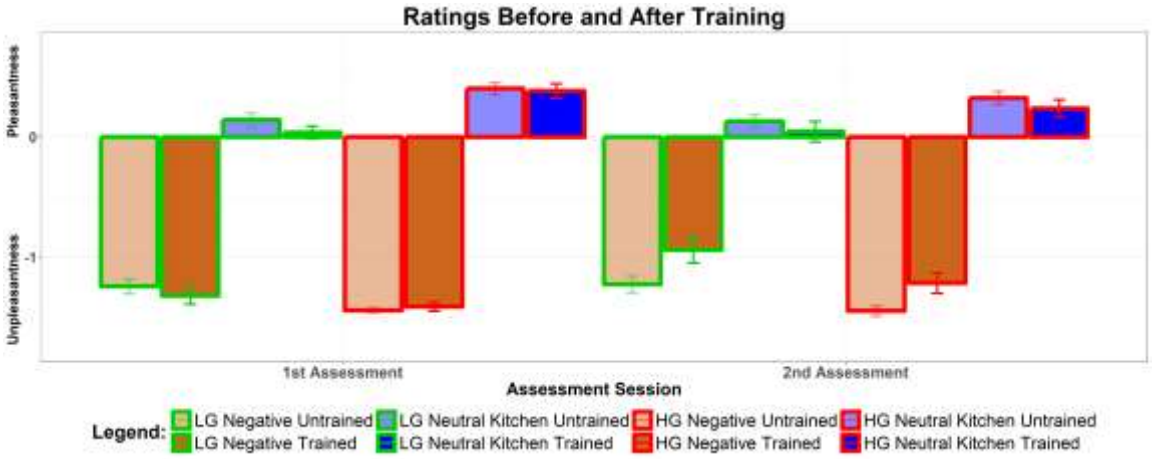


Figure 27: Ratings before and after Training – Average ratings (y-axis) for the negative (brown) and neutral-kitchen (blue) images, pseudo-randomly pre-assigned to be untrained (lighter colour) or trained (darker colour), displayed by the group with low (green outline; LG) and the group with high (red outline; HG) fear of contamination traits at the 1st and 2nd assessments (x-axis). Main results showed that both groups rated the trained negative images as being less pleasant, compared to the untrained ones.

3.4.4 Comparison of Ratings for the Weak versus Strong Negative Images

Results of the MEM analysis for the negative images revealed a main effects for content strength (F(2, 215) = 7.20, p <.001), for which visual inspection of the figure below suggested that it is driven by the fact that the negative strong images were rated as more unpleasant than the negative weak images, in both sessions. Results also showed a main effect of group (F(1, 43) = 15.68, p <.001), for which visual inspection of the figure below suggested that it is driven by the fact that the HG rated both the weak and strong negative images as more unpleasant than the LG. No other significant effects were reported (session: F(1, 215) = 2.59, p = 0.11; IA group*session*content : F(2, 215) = 1.85, p = 0.16).

In addition, the 2-way interaction between content and group at the 1st assessment (Z = -1.90, p = 0.056) was analysed to examine the rating differences between the groups for the weak and strong negative images at the 1st assessment. This would allow to examine if the prerequisites were met, and, consequently, to correctly interpret the analyses afterwards. More specifically, it was expected that the HG would have initially rated the negative strong images as being more unpleasant than the weak negative images, compared to the LG, due to the HG's higher sensitivity for negative stimuli. However, the interaction between content and group was not significant.

Figure 28: Ratings before and after Training – Average ratings (RBs; y-axis) for the negative weak (lighter brown) and negative strong (darker brown) images, by the group with low (green outline; LG) and the group with high (red outline; HG) fear of contamination traits at the 1st and 2nd assessments (x-axis). Main results showed no rating difference in the negative weak versus strong images

3.4.5 Comparison of Ratings for Generalization Assessment

In order to measure the training generalization effects relative to the assessment-only negative images, i.e., to the negative weak and strong images, a correlation was performed with the rating changes (between the two assessments) for the negative weak and strong *versus* the negative medium trained images with the OCI-R scores (see figure 29). Results showed that the OCI-R score did not explain the rating changes in the medium trained *versus* weak ($r = 0.14$, $p = 0.38$) and strong *versus* medium trained ($r = 0.18$, $p = 0.26$) negative images.

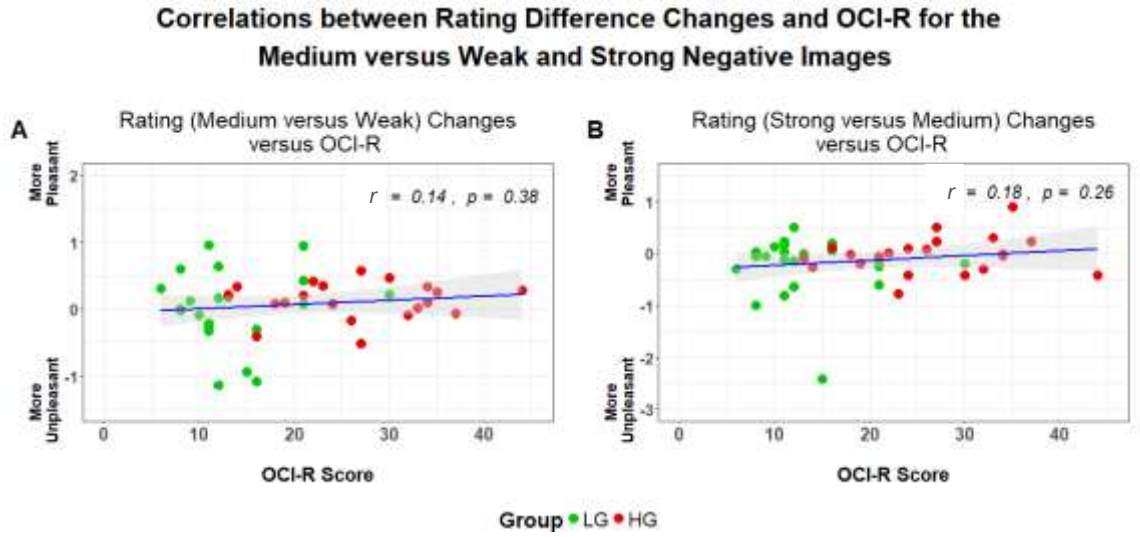


Figure 29: Correlation between Rating Difference Changes and the OCI-R for the Negative Images. **Left Panel** - Average ratings (y-axis) changes (between assessments) for the medium trained versus weak negative images differences by each participant (dots), with their respective OCI-R scores (x-axis). **Right Panel** - Average ratings (y-axis) changes for the strong versus medium trained negative images differences by each participant, with their respective OCI-R scores (x-axis). Red dots – participants from the group with higher fear of contamination traits (HG). Green dots - participants from the group with lower fear of contamination traits (LG).

3.5 Practical Test

In order to analyse training effects for stimuli related to real-life situations, participants performed a final test after the 2nd assessment. This test consisted on two parts, a sitting test and rating of novel images. In the sitting test, the time it took for each participant to pull the chair away from the table, to sit down on a modified pillow and press the spacebar key on the laptop, was measured. Unbeknownst to each participant, for this test the pillow cover on the chair was changed from a completely white cover (which was used during the AAT protocol up until this point) to a cover displaying an image of a dirty toilet. In the ratings performed afterwards, participants had to rate novel negative, positive and neutral images, in terms of their pleasant/unpleasant characteristics.

Regarding the sitting test, it was expected that participants from the HG would take on average more time to sit down on the modified pillow cover, compared to the LG, since the HG would be more sensitive to negative image displayed on the pillow. As for the ratings, it was expected that the HG would rate the negative images as being more unpleasant than the LG, due the same reason as in the previous hypothesis.

In the sitting test, before testing the main hypothesis, prerequisite tests were performed to compare the average time it took for each group to pull the chair away from the table, sit down and press the spacebar *versus* a theoretical zero time. The aim here was to check if on average participants from each group had indeed taken some time to perform the necessary steps before pressing the spacebar, a process which was always accompanied by the instructor. Contrast test results showed that each group took a significant amount of time perform the steps (HG: $Z = 10.9$, $p < 0.001$; LG: $Z = 11.8$, $p < 0.001$). Thus, the prerequisites were fulfilled.

Regarding the main hypothesis, a contrast test was performed to compare the time it took to press the spacebar *between* groups. Before performing this test, a closer inspection of the individual data showed that two individuals in the LG had relatively high values (participant #17: 30 seconds; participant #18: 45 seconds), which seemed to deviate from the rest of the participants. Taking into account these two participants, the contrast test showed no group differences in the time needed to perform the sitting test ($Z = 1.62$, $p = 0.11$; LG: $\bar{x} = 9.65$ seconds, $SD = 9.99$; HG: $\bar{x} = 6.14$ seconds, $SD = 2.02$). However, considering how they seemed to stand-out from the rest of the other participants, the same outlier exclusion criteria used for the RTs in the AAT Assessment and AAT Training (see section *2.4 Data Pre-Processing* in Methods) was used here, upon which the two LG participants were considered outliers and were consequently excluded

for this test. When re-applying the same contrast test, results continued to show no significant group sitting time differences (Z = 0.82, p = 0.41).

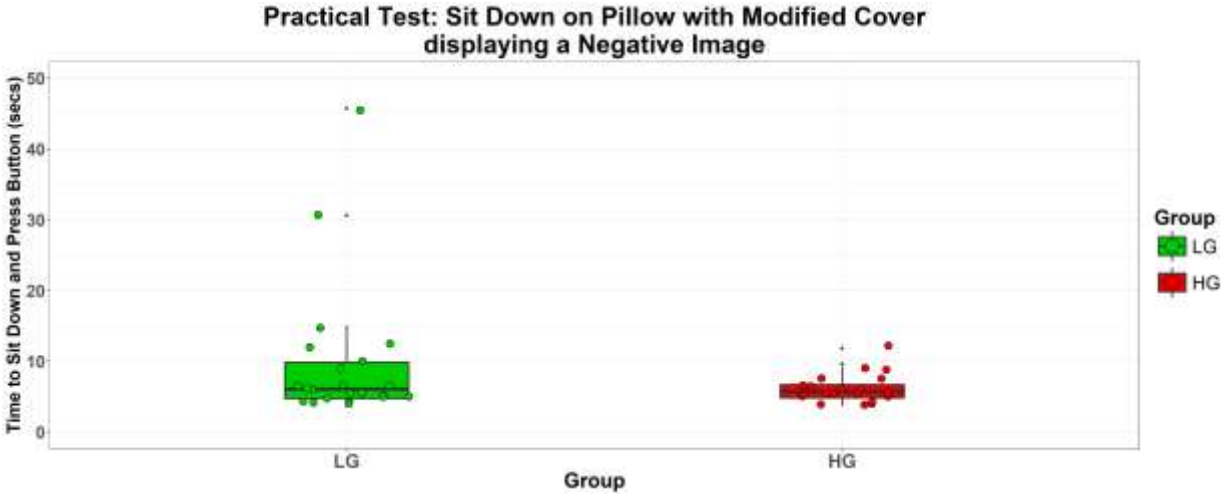


Figure 30: Sitting Test – Average time (in seconds; y-axis) it took participants from each group (x- axis) to pull the chair away from the table, sit down and press the spacebar. Results showed that both groups a substantial amount of time to complete the steps, until the time stopped recording. Each dot represents the average time participant displayed to complete the steps. Green - Group with low fear of contamination traits. Red - Group with high fear of contamination traits.

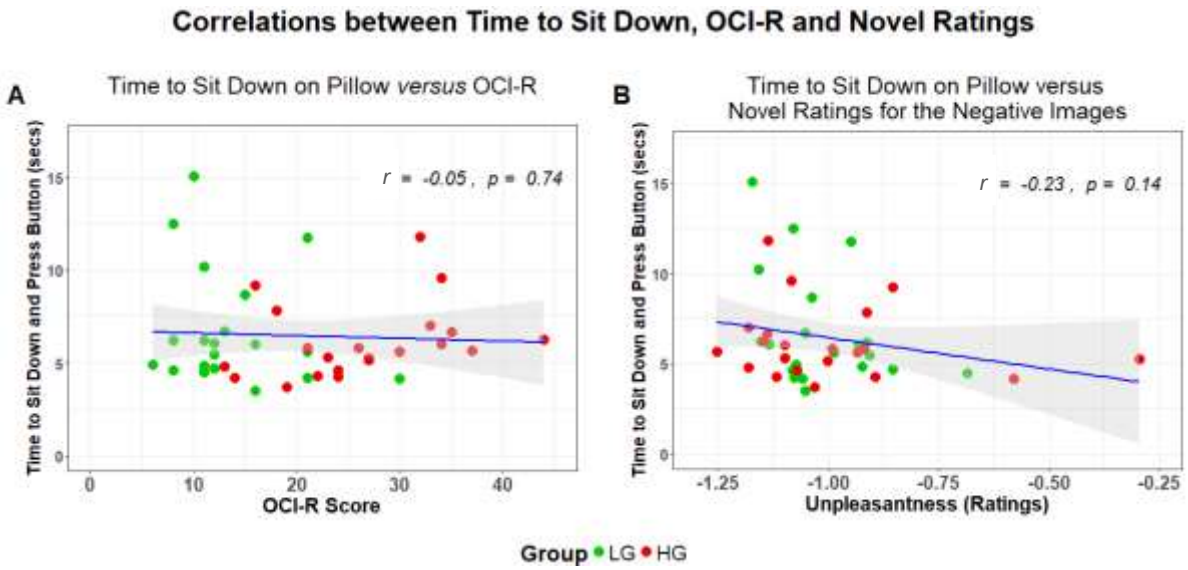


Figure 31: Correlation Between Sitting Time and OCI-R – Average ratings (y-axis) for the negative images of the practical test by each participant (dots), between the 1st to the 2nd assessment, with their respective OCI-R scores (x-axis). This correlation was not significant, indicating that a higher or lower OCI-R score did not explain the time it took to complete the steps. Red dots – participants of the group with high fear of contamination traits. Green dots - participants of the group with low fear of contamination traits.

As an exploratory analysis to check for the influence of individual OCI-R differences, correlations between the time to sit down and press the spacebar *versus* the respective OCI-R scores and the novel ratings, were performed. Both of these analyses revealed non-significant correlations (r = -0.05, p = 0.74), indicating that inter-

individual differences with regard to the OCI-R score did not explain the duration that it took participants to press the spacebar and the rating differences for negative images (r = -0.23, p = 0.14).

In the ratings performed afterwards, results of the MEM reported a main effect of category ($F(2,129) = 932.70$, $p < 0.001$), with no other significant effects (group: $F(1,129) < 0.001$, $p > 0.99$; IA category*group: $F(2,129), = 1.40$, $p = 0.25$). Before testing the main hypothesis, prerequisites tests were performed to compare the ratings for each category to a theoretically zero (neither unpleasant nor pleasant), in each group. The aim here was to check whether each group had perceived the valence of the images as expected: negative images as unpleasant, positive as pleasant and neutral as neither unpleasant nor pleasant. In line with the expectations, both groups rated the negative images as unpleasant (HG: $Z = -22.15$, $p < 0.001$; LG: $Z = -22.54$, $p < 0.001$) and the positive images as pleasant (HG: $Z = 19.73$, $p < 0.001$; LG: $Z = 21.82$, $p < 0.001$). However, while the LG did not have any preference towards the neutral images (LG: $Z = 0.73$, $p = 0.47$), the HG rated them as pleasant (HG: $Z = 2.42$, $p = 0.016$). Thus, the prerequisites tested were closely fulfilled.

Regarding the main hypothesis, in contrast to the expectation, a contrast test showed no group differences in the rating for the negative images ($Z = -0.09$, $p = 0.93$; see figure below). When applying the same contrast tests for the other two categories, results also showed no group differences (positive: $Z = -1.13$, $p = 0.26$, neutral: $Z = 1.22$, $p = 0.22$).

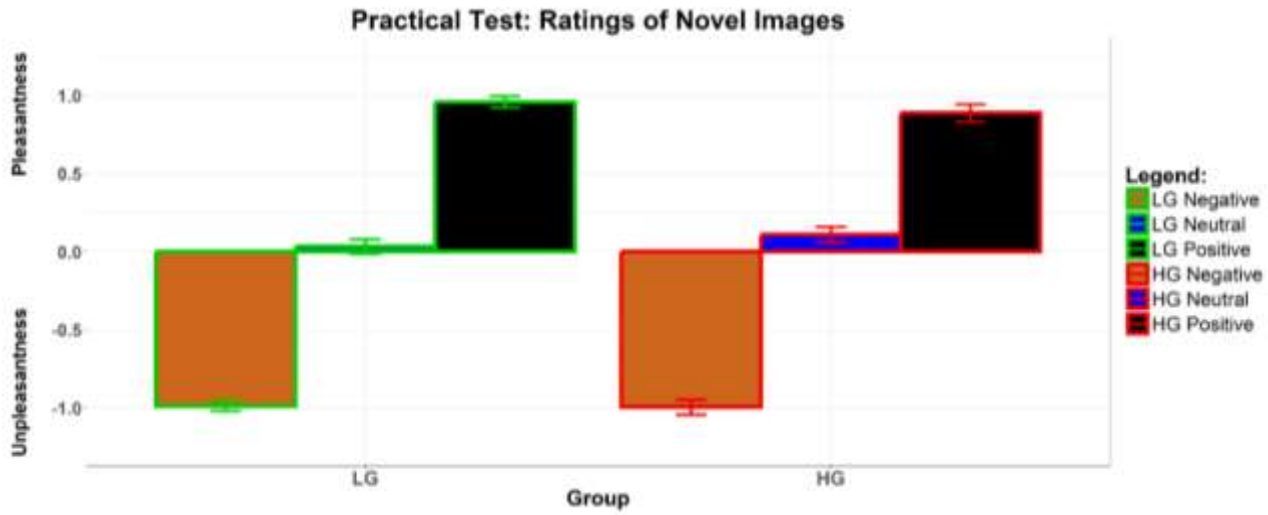


Figure 32: Ratings of the Practical Test – Average ratings (y-axis) for all three category of images (colours), namely positive, negative and neutral, by each group (differed by outline). Main result was obtained by comparing the ratings of the negative images between group, which showed no group differences for this comparison. Red bars – Negative images. Blue bars – Neutral-kitchen images. Green bars – Neutral images. Green Outline –Low fear of contamination traits' group (LG). Red Outline - High fear of contamination traits' group.

4.Discussion

Content:

- ✓ ***Brief Overview of the Current Thesis***
- ✓ ***AAT Training***
- ✓ ***AAT Assessment***
- ✓ ***Practical Test***
- ✓ ***Integration of Training and Assessment Performance***

4.1 Brief Overview of the Current Thesis

Previous research has shown that AAT training protocols have the potential to modify automatic tendencies in individuals with pathologically enhanced automatic tendencies[67] such as in addiction[66,79–83], anxiety-related disorders[125–128] and eating disorders[62–64]. In particular, with regard to OCD, excessive automatic avoidance tendencies have been regarded one of its main symptoms , possibly as a means to avoid contact with distressing stimuli[158–161]. Taking this evidence into account, the current study had the aim of training new behavioural responses towards distressing stimuli, namely automatic approach tendencies to contamination-related imagery, that would contradict the avoidance tendencies.

Upon establishing the aim above, the current thesis assumed that by repeatedly training approach responses towards negative stimuli, new S-R connections (approach the negative stimuli) would get gradually built-up to the point of competition with the old S-R connections (avoid the negative stimuli). If so, the training would have the potential to gradually induce and develop new behavioural habitual response for the negative stimuli. In a clinical context, this could then be used in a way that would help patients to acquire new associations to the feared stimuli, in a less effortful manner. More precisely, this training protocol could be applied as an add-on to the standard ExRPT, in a way that would facilitate the acquisition of new S-R connections through a more implicit procedure, especially in patients who exert a strong degree of cognitive effort to prevent engaging in ritualistic behaviours.

Taking into account this assumption, in the current thesis students were pre-selected according to their fear of contamination traits, that is, whether they had almost none or low fear of contamination (LG) vs above average fear of contamination (relative to a healthy sample, HG), as defined by previous studies[86,153]. Exclusion criteria, namely a history of chronic disease, brain trauma and signs of psychological distress, were assessed to make sure all individuals were healthy.

Once in the laboratory, participants performed an AAT protocol which required them to quickly perform reactions to stimuli presented on a laptop screen, by means of joystick push or pull reactions. In particular, their RBs towards different categories of stimuli, including contamination-related imagery, were assessed before and after a five-day training period. In this training, participants trained to repeatedly perform approach reactions (pull motions) to negative stimuli and to perform avoidance reactions (push motions) to neutral stimuli, the latter as a control condition. Due to the LG's lower sensitivity for contamination-related content, the main initial hypotheses were the following: (1) Throughout training, participants in the LG would speed-up the RTs in the *approach negative* condition more than the HG; (2) Between the assessment sessions, participants in the LG would display a stronger avoidance RB reduction for the negative stimuli, i.e., faster RTs when pulling compared to pushing, compared to the HG.

In line with expectations set for the AAT Training, when approaching contamination-related stimuli throughout the training, the LG speeded-up the RTs of the full joystick movement more than the HG, and more than in the control condition *avoid neutral-kitchen*, whereby these effects were driven by the initiation RTs. Thus, these results indicate that the LG decreased the RTs at the beginning of their approach reactions for the contamination-related stimuli, which suggests that they built-up new S-R connections as initially expected. In contrast, when analysing the end of the joystick reactions, the reversed pattern was found: The HG displayed a stronger decrease of the RTs more than the LG, and more than in the control condition *avoid neutral*, meaning that the HG decreased the RTs for the contamination-related stimuli mostly after initiating their reaction. Importantly, the HG generally displayed faster initiation RTs at the beginning of the training, independent of the condition. This suggests that the HG had enhanced levels of cognitive control throughout the entire training, while trying to end the condition *approach negative* faster than *avoid neutral*. When analysing the approach *versus* avoidance RTs in the AAT assessment, i.e., when participants did not directly pay attention to the content of the stimuli, a general increase of approach tendencies from pre to post-training was observed for the contamination-related stimuli. This effect was driven mostly by the LG than by the HG. In addition, both groups reported that the contamination-related images that were used in the training became less unpleasant from pre to post-training, compared to the untrained contamination-related images (both of medium content strength). These findings are in line with the general assumption that initial conscious awareness is necessary to develop a new habitual response: The consciously reported ratings changed after training specifically for the trained images, while the training period might not have been long enough to already modify reaction tendencies in such a specific way. Further analyses did not reveal evidence for any training effect on the contamination-related images with weak and strong content, which had not been used themselves as stimuli during the training. Here, it is important to note that both groups rated the strong contamination-related stimuli as being more unpleasant than the weak, whereby the HG rated all images in general as being more unpleasant than the LG. Lastly, in the practical test performed at the end of the AAT protocol, no group differences were found in the time to sit down on the pillow displaying a contamination-related image.

4.2 AAT Training

In the AAT training version, participants repeatedly performed two conditions: *approach negative* and *avoid neutral-kitchen*, where they had to approach contamination-related images and avoid kitchen-related images, respectively. The former was meant to induce a gradual acquisition of new automatic (approach) response tendencies towards the negative stimuli, while the latter served as the control

condition. Overall, results showed that the groups did not differ at the intercept, but – as expected – both displayed a gradual decrease in the RTs in the approach negative and avoid neutral-kitchen condition

In particular, the LG displayed a steeper learning curve for the *approach negative* condition compared to the *avoid neutral-kitchen*, and compared to the approach negative in the HG. Before interpreting all the results in a more specific manner and in light of previous evidence, it is important to note that to the author's knowledge, this is one of the first studies, alongside[129,162], to analyse RTs and ratings *throughout* the training period instead of only assessing implicit and explicit behaviour before and after the training period.

4.2.1 No Group RT Differences at the Intercept

With regard to the intercept of the training curves, it was expected that the different fear of contamination traits between groups would be reflected in higher RTs for the *approach negative* condition in the HG, due to their higher sensitivity for contamination-related images, compared to the LG. In contrast to this expectation, results showed no group RT differences for the *approach negative* condition at the intercept, which hints to a possible cognitive control involvement throughout the training. In fact, this could have "masked" intercept group differences in the power-law fitted curves for the *approach negative* condition. Nevertheless, the similar intercepts that both groups displayed for this condition facilitate the interpretation (below) of RT changes within and between groups along the training, when analysing the slopes of the *approach negative* and *avoid neutral-kitchen* conditions.

4.2.2 Overall Slower Full Joystick Movements in the HG

In line with the expectations, results showed that the groups differed in the slopes for the *approach negative* condition, whereby the LG performed the *approach negative* condition progressively faster, more than the HG. This group difference can be explained if one assumes that the HG exhibited relatively stronger cognitive effort due to their stronger difficulty while repeatedly approaching contamination-related images, given their higher fear of contamination traits. In addition, such increased cognitive effort might also be responsible for the fact that the HG did not demonstrate different RTs speeding-up between the *approach negative* and the *avoid neutral-kitchen* conditions during training, as opposed to the LG.

Accordingly, previous literature indicates that cognitive control is needed to appropriately adjust goal-directed behaviour in a flexible manner[163,164]. In fact, one of the executive functions that requires cognitive involvement, namely flexibility, is response inhibition, which is required in situations when the automatic response needs to be inhibited and replaced by a new response, to adapt to a new context[163]. As such, these findings (described in the paragraph above) support the idea of a general cognitive involvement in the training specially in the HG. More precisely, due to their higher fear of contamination traits, the HG might have paid increased attention to inhibit their

automatic tendencies and correctly perform the response associated with each stimuli, particularly when performing incongruent responses to contamination-related images. In fact, previous studies have reported cognitive inflexibility in OCD patients and in healthy students with OCD-like traits[165–167], which supports the idea of an increased cognitive effort by the HG mentioned above. However, it is important to note that the training design used here was still easy enough to for participants with fear of contamination traits to perform, as shown by the small number of error-trials throughout the training period.

4.2.3 Groups Differed at Separate Subcomponents of Joystick Movement

To the author's knowledge the present work is the first study to analyse subcomponents of the joystick movement *during* training. To briefly remind the reader, the initiation RTs refer to the time it took to *tilt* the joystick in each trial, whereas the motion RTs refer to the time from the joystick tilt until the joystick reached the *end position*, i.e., the movement was completed, in each trial.

Results of the joystick movement subcomponents' analyses indicate that the LG progressively speeded-up their joystick tilting movements in the *approach negative*, compared to the *avoid neutral-kitchen* condition, and compared to the HG in the *approach negative* condition. Given their lower fear of contamination traits and possibly less cognitive effort, these results point to the LG's increasing ability to quickly recognize each stimulus (negative / neutral) with the respective reaction (approach / avoid). Interestingly, the HG was faster at tilting the joystick for both conditions, but were significantly slower in speeding-up throughout the training, compared to the LG. This suggests that the HG displayed a constant baseline attentional effort during training, as discussed in the previous section, which then could have leaked in the joystick reaction movements. This is particularly shown by the HG's speeding-up of joystick completion movements in the *approach negative*, more than in the *avoid neutral-kitchen* condition – whereby such difference was not observed in the full joystick reactions – and more than the LG in the *approach negative* condition. As such, following up on the reasoning above, the higher attention levels the HG seems to have displayed could have enabled them into exerting additional conscious control specifically to quickly end the incongruent response, i.e., approaching contamination-related images, which improved during training.

Evidence supporting cognitive control involvement in the latter portion of the joystick movement, as opposed to the beginning, comes from studies like Sheridan *et al.*[168], which analysed the initial reaction time and movement time (equivalent to the initiation and motion RTs in this thesis, respectively) upon moving a vertical level in response to a stimulus presentation on a laptop screen. This study, alongside others[168,169], suggested that the *initiation* RTs were a reflection of the time required to select an appropriate response, but more importantly, that the *motion* RTs could be visually guided *if* sufficiently long enough to involve additional cognitive control[168,169]. In fact, the AAT Training used here required participants to first recognize the image as

negative or neutral and only then perform the correspondent action, rather than just react as fast as possible to push a button. As such, one can argue that the AAT has more common features with a choice reaction-task (stimuli are paired with different motor responses) than with simple reaction-tasks (only one motor response is possible)[170] used in the aforementioned studies. Thus, given that the initiation RTs values obtained here were much higher than the motion RTs, these studies provide support for the fact that the HG was able to partially control the latter portion of the joystick movement. More precisely, since the HG had higher fear of contamination traits, they could have exerted additional cognitive in order to complete the approach response for the contamination-related stimuli as fast as possible.

Thus, it is recommended that future AAT training studies should also analyse these RTs subcomponents in order to assess how exactly differences in disorder-related symptom severity are reflected in the RTs. If so, one non-invasive methodology that should be considered for this is electroencephalography (EEG) that allows to measure brain activity in the form of event-related potentials (ERPs). In short, ERPs reflect the summed activity of synaptic potentials when large numbers of neurons fire synchronously while processing relevant information, which can be analysed in a simple EEG recording, allowing to study physiological correlates of mental processes. More precisely, ERPs might help to further distinguish between automatic and cognitively-mediated processes associated with the joystick movements. Indeed, there is evidence showing that changes in ERPs may provide useful information about disorder-related symptoms[171,172] and for evaluating pharmacological and therapy outcomes[173] in OCD. As such, it would be interesting to investigate how ERP markers in OCD would correlate with the joystick movement differences.

4.3 AAT Assessment

In the AAT assessment, participants were instructed to approach and avoid images containing either positively, negatively or neutrally-valenced content. However, instead of paying attention to the content of the images as in the training, participants were instructed to approach or avoid these images depending on the direction of an arrow presented alongside each image. In addition, participants had to rate the content of the images in terms of their pleasant/unpleasant characteristics. With this framework, participants' RB and rating changes between the 1st and 2nd assessment were analysed, particularly for the contamination-related stimuli, search for training effects.

4.3.1 RBs and Ratings at the First Assessment before Training

Concerning the RBs, the majority of prerequisites tested were not fulfilled. Here, the expectations that both groups would display an avoidance RB for the negative images, whereby the HG would have a stronger baseline avoidance RB than the LG due

to higher fear of contamination traits, were not met. In fact, not only there were no group differences, the HG did not display any bias and the LG actually displayed a significant approach RB for the negative images. The latter was consisted with the general motor approach bias this group also displayed in the AAT Arrow, i.e., when reacting to the presentation of arrows alone (no additional images shown). Such finding could be attributed, perhaps, participants' personality traits (e.g., more pronounced exploratory behaviour in the LG than HG) and future analyses on the self-report questionnaires will shed light on this matter. No other significant RB were obtained for the neutral-kitchen, neutral-street and positive images. Nevertheless, the analyses on the RBs changes between the 1st and the 2nd assessment were interpretable, since they inherently took into account the RBs displayed at baseline.

With regard to the ratings, all prerequisites were fulfilled, whereby results showed that both groups rated the negative images as being unpleasant, the positive images as being pleasant and the neutral-street images as neither pleasant nor unpleasant, as expected. As for the neutral-kitchen, however, both groups rated them as being pleasant, although visual comparison with the negative and positive suggest that they were still considered relatively neutral. In addition, in line with the initial expectations, results revealed that the HG rated the negative images as being more unpleasant than the LG, which can be attributed to the HG's higher fear of contamination traits. Moreover, the HG rated the positive images as being less pleasant and the neutral-kitchen images as more pleasant, than the LG. No group differences were found for the neutral-street images.

4.3.2 Stronger Training Effects in the Ratings than in the Reaction Biases

The AAT protocol used in the current thesis allowed to evaluate if the ratings and the RBs displayed in the AAT assessment for the trained images, i.e., the images that were also shown in the AAT Training, would change between the assessment sessions as a result of training, compared to the untrained images.

Results revealed a significant, although unspecific, training effect in the RBs displayed for all negative images – of medium content - which was found to be driven by an approach RB increase, between the 1st and the 2nd assessment, in the LG more than the HG. Results also showed a *marginal* approach RB increase between assessments *specifically* for the trained negative images, compared to the untrained negative ones. However, this trend was also observed for the trained neutral-kitchen images, compared to the untrained neutral-kitchen images. Therefore, despite providing a first hint for a change in implicit behaviour, these results do not fully match the hypothesis: a significant *avoidance* RB decrease was expected specifically for the negative trained images more in the LG than in the HG. However, the results obtained can be partially explained in light of the RBs displayed at the 1st assessment, where the LG already displayed an approach RB for the negative images and whereupon the training could have enhanced that tendency.

With regard to the ratings, in line with the expectations, results indicate training effects for the trained stimuli, whereby both the trained negative and neutral-kitchen images were rated by both groups as being significantly less unpleasant from the 1st to the 2nd assessment, compared to the untrained negative and neutral-kitchen images. In fact, interpretability of these training effects is further supported by the fact that there were no differences between the trained and the untrained images at the 1st assessment (before training), both in the negative and neutral-kitchen images.

Overall, results indicate stronger training effects in the ratings than in the RBs, as shown by the significant unpleasantness decrease for the trained negative images *versus* the marginal approach RB increase for the trained negative images, compared to the untrained ones, in both groups across assessments. In the context of the current framework, this is in line with the neurobiological mechanisms involved in the formation of new habitual behaviours, namely in the building-up of new S-R connections[8,9]. More precisely, in the process of acquiring a new habitual response, before new S-R connections strengthen enough to the point of automaticity, performing new actions requires a more conscious and controlled response[8,9]. Indeed, this may be why, in general, participants reported an explicit evaluation change towards the contamination-related images, as opposed to displaying RB changes, which reflect implicit behaviour. Moreover, if the training had been prolonged enough, the explicit evaluation changes could have led to significant implicit behaviour changes as a result of incremental S-R strengthening. In fact, the *marginal* RB changes across both groups for the trained negative images between the 1st and the 2nd assessment might hint to that.

With regard to previous AAT training studies that applied explicit and implicit measures, the results above are in line with two studies that reported explicit behaviour changes, but no implicit changes[58,59]: in children instructed to repeatedly approach or avoid novel animals, there was an increase in self-reported liking of the approached animal and in the self-reported disliking of the avoided animal, together with an explicit higher fear for the avoided animal. Interestingly, in both of these studies, there was no significant difference in implicit attitudes after training, i.e., in how fast children associated each animal with either a positive or negative label[58,59]. One study, however, reported the opposite pattern, although less pronounced[60]: subjects were trained to approach or avoid images with different emotional expressions depending on their background colour. Before and after training, they had to correctly categorize negative or positive words as pleasant or unpleasant, as fast as possible, and then to rate all emotional faces based on several traits such as attractiveness and friendliness. More importantly, in the categorisation task, before a word appeared on the screen, a priming image (that had been either approached or avoided in the training) was shown for 300ms. Results showed no changes in the explicit ratings and in (implicit) categorization evaluations for angry and smiling faces, whereupon the only training effect was found in the implicit evaluation for the neutral faces. Here, the authors argued that the training was too short and/or weak to affect the strong valence of the angry and smiling faces,

as opposed to the ambivalent nature of the neutral stimuli[60]. However, such pattern was not found in the current work, since results indicate stronger training effects for the contamination-related images compared to the neutral images.

Furthermore, a significant interaction between the trained/untrained images, assessment sessions and groups was expected. More precisely, it was theorized that the LG would more easily change their pre-existing bias to avoid negative stimuli after training, compared to the HG that was more sensitive to this content, due to the HG's hypothesized stronger automatic avoidance tendencies. However, this expectation was not met since the 3-way interaction was not significant, neither for the RBs nor the ratings. As such, the results obtained do not fully support the initial hypotheses. In fact, the results are in contrast with previous studies who showed a clear RB change specifically for the trained stimuli, in individuals with obesity[174], with heavy alcohol drinking behaviour[79–82] and with fear of contamination traits[127]. However, any comparison with these studies needs to be performed with caution, since only one assessed participants' baseline RBs before training *and* reported a significant RB change[79]. Nonetheless, the results of the current work are comparable to the aforementioned studies, in the sense that training effects were also analysed by comparing the changes between the 1st and the 2nd assessment, and, hence, taking into account the RBs variance displayed at baseline.

4.3.3 Stronger Reactions for the Most Unpleasant Negative Images

To the author's knowledge, this is the first study to analyse the response of participants with fear of contamination traits to different degrees of the feared stimuli. In the current study, this was performed by analysing the RB and the ratings displayed for mild (weak) and exacerbated (strong) contamination-related images (note: the results for the images with medium content were already discussed before). The hypotheses established here were that the HG would display stronger RBs and unpleasantness feelings for the strong negative images than for the weak negative images, more than the LG, due the HG's higher fear of contamination traits.

With regard to the RBs, in contrast to the expectations, results showed that both groups displayed a *marginal* stronger approach RBs for the strong negative images, compared to the weak ones, in the two assessment sessions. Moreover, the LG *marginally* increased their approach RBs across the two assessment sessions, compared to the HG. Concerning the ratings, in line with the expectations, both groups rated the strong negative images as being significantly more unpleasant than the weak ones in both assessments, whereby the HG rated the former as being *marginally* more unpleasant than the latter, compared to the LG. Thus, results indicate that participants displayed a tendency for stronger *approach* reactions for the most *unpleasant* negative images, which at first glance may look like contradictory evidence. However, if a form of cognitive control is assumed to be involved, then one can argue that these results indicate that participants put more of such cognitive effort in the joystick movement

reactions for the strong negative stimuli. Additionally, the results further support the notion of an increased cognitive awareness already discussed above, as seen here by the stronger changes in explicit behaviour, captured by the ratings, than in implicit behaviour, captured by the RBs.

It is important to also note that in a clinical context, the evaluation of implicit and explicit responses for different degrees of the fearful, disorder-relevant stimuli is important in the selection of the stimuli used for the ExRPT: for the therapy to be successful, throughout the initial sessions, the degree of exposure must be demanding enough to activate feelings of fear, while being low enough to allow for gradual habituation processes to occur. As the patient progresses, later sessions use more demanding stimuli. Thus, the analysis of participants' reactions for different degrees of the fearful stimuli performed in this study provides a first step for future studies to select adequate stimuli to be used in the ExRPT and the AAT add-on training, so that patients experience less cognitive effort and consequently have less probability to drop-out.

4.3.4 No Hints for Generalization Effects

With regards to possible generalization effects, the design used here allowed to evaluate if training effects would also be observed for untrained stimuli, specifically for the negative medium untrained and weak images (shown only in the assessment sessions). The expectation was that as a result of the AAT Training and the theorized S-R connections strengthening, the trained negative images – that were of medium content strength – would have elicited a similar avoidance RB and unpleasantness rating as the weak negative images at the 2nd, but not yet at the 1st assessment. In contrast to the expectation: the correlations between participants' OCI-R scores with the differences across assessments (1) between the strong and trained negative images and (2) between the weak and trained negative images, showed that the inter-individual differences in the OCI-R scores had no impact in generalization effects.

Regarding previous studies, so far only a reported successful generalization of training effects to untrained stimuli, namely in heavy drinking students[80,82] and patients with alcohol addiction[79]. To the author's knowledge, only one non-alcohol-related study has investigated generalization effects after training[174]. In this study, despite of successfully decreasing an approach RB for unhealthy food in individuals with obesity, the training effects did not generalize for untrained unhealthy food images[174].

Taken together, it is recommended that future studies should emphasize the sensitivity of the mild and exacerbated degrees of the feared stimuli, for example using clean toilets and excessively dirty toilets, respectively, in order to better test the generalization capability of the training protocols. Moreover, a sample of OCD patients might be better to study this, due to their more severe obsessive-compulsive symptoms, as opposed to just fear of contamination traits.

4.4 No Group Differences in the Practical Test

After performing the 2nd assessment, all subjects underwent a test to assess their avoidance tendencies for a contamination-related stimulus that was closer to a real-life situation than the AAT. This test consisted in measuring the time it took for each participant to pull a chair away from the table, sit down on a modified pillow – which displayed a contamination-related image - and press the spacebar on the laptop. This allowed to investigate if the sitting time was related to participant's fear of contamination traits. Afterwards, participants rated novel images, which contained either negative, positive or neutral characteristics, in order to investigate whether sitting on the modified pillow had an influence of participant's general mood. Respectively, the expectations were that the HG would take more time to sit down on a pillow displaying a negative image, and that this would influence the HG to rate the novel negative images as being more unpleasant, compared to the LG.

In contrast to the expectations, results did not show any significant group differences neither in the sitting time nor in the novel ratings. Nonetheless, in these ratings, each group rated the negative as being unpleasant, the positive as being significantly pleasant, as expected. Here the only difference was reported for the neutral images, whereby the LG rated them as being neither unpleasant or pleasant, the HG rated them as slightly pleasant.

Taking into account the results above, it is important to refer that the design of the practical test used in the current thesis was developed in a previous study in the lab[129] and its design remained unchanged. Compared to other studies who reported changes in the behaviour also in the practical test, the design used here might have been less realistic. The high variance in the sitting time in the current study does hint to that. In contrast, previous studies evaluated participants' behaviour to real-life disorder-related stimuli. One example of this is the test used by Wiers and colleagues[80], where they asked participants to drink three colas and three beverages in order to guess and rate the brands. With this design, results showed that participants who trained to avoid alcohol drank less beer in this taste test. Other studies, although not using specific tests after training, evaluated specific disorder-related measures such as relapse rates in individuals with alcohol addiction and cigarette/nicotine consumption[66,79]. One important note is that comparison to these studies needs to performed with caution, since they assessed the aforementioned measures not directly after training but rather at a follow-up, i.e., with a specific time window after training. Thus, future work should go into improving the procedure when sitting down in the chair displaying contamination-related stimuli, in order to reduce noise in the data. Alternatively, a more robust practical test could be developed for the current AAT protocol, that takes into

account its sensitivity for OCD-related traits and the variance inherent to the task, to also reduce variance in the data, and what participants/patients what comfortable or willing to do.

4.5 Integration of Training and Assessment Performance with Explicit Ratings

4.5.1 Conscious Re-Appraisal of Negative Stimuli

Results are in line with literature on habit formation, particularly on the transition of goal-directed behaviours to habitual responses and the inherent building-up of new S-R connections. Accordingly, in order to build a new habit, one initially needs a high degree of conscious awareness and effort to perform the new action, which, if continuously repeated, leads to a gradual increase in behaviour automaticity, by incremental strengthening of S-R connections[8,9]. In the current study, this is evidenced by the gradual RTs speeding-up for the *approach negative* during training and by the stronger training effects captured explicitly than implicitly in the assessment: Across both groups, the trained negative images were rated as becoming less unpleasant, whereas they elicited only a *marginal* increase in the approach RB, from the 1st to the 2nd assessment, compared to the untrained negative stimuli. Thus, the training period might have been relatively adequate to evoke subtle conscious re-appraisal changes, but not long or effective enough to observe significant implicit behaviour changes after training for the trained negative stimuli, when not directly paying attention to the content of the stimuli.

In addition, the results on the implicit and explicit behaviour changes mentioned above corroborate the potential clinical application of an AAT training protocol in OCD patients, while undergoing the effortful procedure of ExRPT. More precisely, the AAT could facilitate the process of acquiring new behavioural appropriate (approach) responses in the ExRPT in a more implicit and less effortful manner, due to its computerized nature compared to the exposure characteristics of the ExRPT. Consequently, this could lead to the building-up of new S-R connections in a way that the new responses could persisted even situations with depleted cognitive control. Thereby, the implementation of this training via novel and flexible technologies, such as a smartphone app, could provide a realistic framework for regular, daily training sessions (for a similar application see [175]).

4.5.2 Hints for Cognitive Control

Results point to a possible cognitive control involvement in both groups throughout the protocol, especially during training due to a possible conscious effort in

performing incongruent actions, i.e., approaching contamination-related images. This could explain why neither groups performed the *avoid neutral-kitchen* faster than the *approach negative* condition, since, in theory, performing an action with a neutrally-valenced image should have been easier (lower RTs) than performing an incongruent response.

In particular, the HG seemed to have displayed a relatively increased cognitive effort throughout the training, possibly reflecting their increased difficulty in approaching contamination-related images due to their higher fear of contamination traits, which was evidenced by: (1) The HG was faster in the tilting joystick movements for both, the negative and neutral stimuli, compared to the LG. However, compared to the LG, the HG showed less decrease of this movement part for approaching negative images and no differences to the neutral conditions; (2) The HG speeded-up the motion times of the *approach negative* condition more than the LG and more than for *avoid neutral*, possibly to consciously end the incongruent response quickly; (3) Strong cognitive control might also have led the HG to not perceive the contamination-related images more negative after the first training session than the LG, while such a group difference was reported in the assessment ratings before the first training. Moreover, there was also no influence of time on the ratings directly after the trainings.

As for the LG, in the assessment version, participants showed no avoidance bias for negative images, a finding that is similar to previous findings[81,83,125,126,157,176] . Rather, they showed an approach RB at the 1st assessment. Furthermore, the LG also showed an increase of this approach RB from the 1st to the 2nd assessment and a stronger approach RBs increase for the most unpleasant negative images, compared to the HG. In how far these findings can be interpreted as signs for 'hyper-regulatory' attention / cognitive control – as suggested by previous studies[177] – will be a subject of future analysis steps. In the current study, the LG showed a general approach bias across categories and – given the group differences in the subcomponents of the training RTs – also splitting the assessment RTs into initiation vs motion times might improve the understanding of these effects.

One might also argue that the ratings used in this study might not have been ideally designed to better capture a conscious re-appraisal of the trained contamination-related images, both in the training and in the assessment. Thus, it is recommended that future AAT training studies improve the way how participants explicitly rate the images throughout the protocol, for example by instructing participants to rate images with regard to multiple dimensions of negative emotions, such as fear *versus* disgust, to better capture explicit changes regarding specific OCD-like traits.

4.6 Limitations

A possible limitation that could have influenced the overall findings is that the fear of contamination traits between groups might not have been different enough in a way that significant group differences would be detectable in the RTs. Pre-selecting participants with higher washing scores could have resulted in more significant group differences in the assessment. However, it is important to note that pre-selection of individuals with washing scores close to its maximum value was difficult, given the fact that this study aimed at only including healthy individuals without any pronounced mental problems. Previous studies performed in healthy subjects have used more elaborate inclusion criteria such as: (1) Using a cut-off of an inventory (Maudsley Obsessional Compulsive Inventory) which was 2 standard deviations above the mean for the normal population. This could have helped to recruit participants whose OCD traits differed significantly more than the average population[127], as opposed to the cut-off used in this study which was based on the washing scale mean that OCD patients displayed in the OCI-R[86,153]; and (2) Objective disorder-related behaviour, such as frequency of alcohol drinking[80,83,157]. Thus, future AAT protocol designs might benefit from more specific inclusion criteria similar to the aforementioned studies, in order to target specific psychiatric traits. However, it is important to note that in the case of current study the pre-selection process found very few participants with a maximum score in the washing scale, as stated above. Therefore, even if such criteria had been applied to this study, its contribution to the overall results would have been most likely non-significant.

In addition, the use of RBs to analyse training effects at the level of implicit behaviour might have hidden weak RT changes in just one of the two directions the AAT Assessment. In fact, one study[78] assessed approach-avoidance tendencies to contamination-related images in students with high obsessive-compulsive (HC) and low obsessive-compulsive (LC) symptoms. Interestingly, whereas the HC group showed a slower approach of contamination-related than neutral pictures as expected, this group did not push away the negative stimuli faster than the LC. Here, the authors suggested that individuals with high OC traits may not necessarily display stronger avoidance tendencies but rather an impaired inhibition of these tendencies[78]. However, one important advantage of using the RBs in the analysis is the better control of within-subject variance. Nevertheless, further analyses of the current data will look at the single directions.

Moreover, the hints to cognitive control – already discussed above - displayed by both groups, particularly in the joystick reactions, could have prevented the AAT protocol to directly target participants' fear of contamination avoidance traits. If such a protocol was to be used in OCD patients with very strong contamination traits and less ability to cognitively influence their reactions, then one could assume that their biases

to contamination-related images were more clearly reflected in the implicit behavioural measure.

Given the goal of recruiting healthy participants with fear of contamination traits from a students' population, two mechanisms were used to ensure participants' mental health. First, participants were explicitly asked whether they had any history of psychiatric or neurological disorders, as well as any type of chronic disorders (see section *2.3 Sample Description* in the Methods). Second, a frequently used screening questionnaire, the BSI, was applied to compare participants' current level of mental distress to an established cut-off. Participants were not invited to the study if they scored above this cut-off; they were advised to contact health services. A full clinical interview according to DSM-V criteria was considered to not be adequate for the current study design, since a large number of potential participants needed to be screened (N≅300, according to the calculations sent and approved by the Ethics Committee of the Medical Academic Centre in Lisbon; Ref.: 307/17). Such an elaborated and lengthy interview, using the Yale-Brown Obsessive Compulsive Scale, would not be appropriate in the context of the current thesis, which focused on healthy subjects with fear of contamination traits. This is because the scales' psychometric properties (sensitivity and specificity) are targeted for individuals with clinically significant OCD symptoms[178–180]. In this perspective, such an interview would rather be the next step in the assessment to give a diagnosis and to tailor treatment after a participant having surpassed the cut-off in the BSI. It is important to note here that previous AAT training studies in the context of OCD such as Amir *et al.*[127] and Najmi *et al.*[78] just obtained their samples "*…from a pool of undergraduate students at a large university...*" (original citations, pages 3 and 4 of the journal articles, respectively) and did not specify exclusion criteria. However, to our opinion, it is necessary to screen samples of young adults, since this part of the general population is definitively not free from mental disorders[181].

Relying on a self-report questionnaire makes the current study susceptible for the general limitations of this method: (1) It may evoke a sense of social acceptance, leading to omission of information due to embarrassment in describing personal problems or insecurities; (2) A lack of objective hindsight can lead to poor self-judgement, resulting in erroneous information from the subjects. Yet, a clinical interview also heavily relies on the verbal answers of the patient, while the interviewer can only partly interpret additional features such as facial expressions. Potential participants for our study knew right from the beginning that they would have to come to the lab several times and be in personal contact with the experimenter. 24 of such 343 participants (7%), who filled in the screening, got BSI scores above the cut-off. This number is comparable to studies that previously used the BSI[112]. Therefore, we assume that participants followed our instructions and chose their answers as close as possible to their real situation.

4.7 Future Plans

With the signs of cognitive effort discussed above, future work will focus on analysing and quantifying the cognitive involvement displayed in the AAT Assessment. More precisely, the initiation and motion subcomponents of the joystick movement will be analysed to investigate if and how an increased cognitive control leaked onto the reactions, when not directly paying attention to the stimuli itself. Moreover, it will also help to investigate whether this possible cognitive involvement differed between groups and between the assessment sessions, due to, respectively, different fear of contamination traits and training effects.

In addition, future analyses will involve novel computational models, previously developed in the lab, to capture the influences of automatic and cognitive processes in the collected behavioural training data. This will be done by using information such as actions performed given the instructions, stimuli characteristics and ratings for each participant. In particular, given the model parameters which include a cognitive control component, this will allow to further analyse, in a different perspective, participants' ability to guide actions in incongruent responses.

Lastly, with regards to the OCI-R and BSI questionnaires, as mentioned in the section *2.2.3 Questionnaires* in the Methods, the translation of all questionnaires was performed according to previously established guidelines[141]. Accordingly, with regards to the OCI-R and BSI scales, when comparing the CFA analyses and internal consistency between the versions used in this thesis and the previous published versions (see attachments section *6.4 Psychometric Analyses of the OCI-R and BSI*), the results demonstrate that the quality of the versions used are comparable to both the adapted Portuguese and original English versions. Nevertheless, using the published Portuguese versions with the current protocol is of interest for future work, in order to compare results more easily to any other study that uses the OCI-R in the Portuguese population.

5. References

1. Evans, J. S. B. T. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* **59**, 255–78 (2008).

2. Strack, F. & Deutsch, R. Reflective and impulsive determinants of social behavior. *Pers. Soc. Psychol. Rev.* **8**, 220–247 (2004).

3. Baumeister, R. F., Bratslavsky, E., Muraven, M. & Tice, D. M. Ego depletion: Is the active self a limited resource? *J. Pers. Soc. Psychol.* **74**, 1252–1265 (1998).

4. Pacherie, E. The role of emotions in the explanation of action. *Eur. Rev. Philos.* **5**, 53–92 (2002).

5. Frijda, N. H., Ridderinkhof, K. R. & Rietveld, E. Impulsive action: emotional impulses and their control. *Front. Psychol.* **5**, 518 (2014).

6. St Quinton, T. & Brunton, J. A. Implicit processes, self-regulation, and interventions for behavior change. *Frontiers in Psychology* **8**, 5 (2017).

7. Gardner, B. A review and analysis of the use of 'habit' in understanding, predicting and influencing health-related behaviour. *Health Psychol. Rev.* **9**, 277–295 (2015).

8. Gasbarri, A., Pompili, A., Packard, M. G. & Tomaz, C. Habit learning and memory in mammals: Behavioral and neural characteristics. *Neurobiol. Learn. Mem.* **114**, 198–208 (2014).

9. Graybiel, A. M. Habits, Rituals, and the Evaluative Brain. *Annu. Rev. Neurosci.* **31**, 359–387 (2008).

10. Neal, D. T., Wood, W. & Quinn, J. M. Habits - A repeat performance. *Curr. Dir. Psychol. Sci.* **15**, 198–202 (2006).

11. Wood, W. & Neal, D. T. A New Look at Habits and the Habit-Goal Interface. *Psychol. Rev.* **114**, 843–863 (2007).

12. Andrew J. Elliot1, 3 and Martin V. Covington. Approach and Avoidance Motivation Across Cultures. in *Handbook of Approach and Avoidance Motivation* doi:10.4324/9780203888148.ch33

13. Eastwood, J. D., Smilek, D. & Merikle, P. M. Differential attentional guidance by unattended faces expressing positive and negative emotion. *Percept. Psychophys.* **63**, 1004–1013 (2001).

14. Solarz, A. K. Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *J. Exp. Psychol.* **59**, 239–245 (1960).

15. Ohman, A. Face the beast and fear the face: Animal and social fears as prototypes for evolutionary analyses of emotion. *Psychophysiology* **23**, 123–145 (1986).

16. Smith, C. A. & Ellsworth, P. C. Patterns of cognitive appraisal in emotion. *J. Pers. Soc. Psychol.* **48**, 813–38 (1985).

17. Bargh, J. A. & Williams, E. L. The Automaticity of Social Life. *Curr. Dir. Psychol. Sci.* **15**, 1–4 (2006).

18. Phaf, R. H. *et al.* Approach, avoidance, and affect: a meta-analysis of approach-

avoidance tendencies in manual reaction time tasks. (2014). doi:10.3389/fpsyg.2014.00378

19. Davidson, R. J., Ekman, P., Saron, C. D., Senulis, J. A. & Friesen, W. V. Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology. I. *J. Pers. Soc. Psychol.* **58**, 330–41 (1990).

20. Lang, P. J., Bradley, M. M. & Cuthbert, B. N. Emotion, attention, and the startle reflex. *Psychol. Rev.* **97**, 377–95 (1990).

21. Chen, M. & Bargh, J. A. Consequences of Automatic Evaluation: Immediate Behavioral Predispositions to Approach or Avoid the Stimulus. *Development* **25**, 1251 (1999).

22. Cacioppo, J. T., Priester, J. R. & Berntson, G. G. Rudimentary Determinants of Attitudes. II: Arm Flexion and Extension Have Differential Effects on Attitudes. *J. Pers. Soc. Psychol.* **65**, 5–17 (1993).

23. Schnerila, T. C. An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal. - PsycNET. (1959). Available at: https://psycnet.apa.org/record/1960-05385-003. (Accessed: 10th May 2019)

24. Zajonc, R. B. On the primacy of affect. *Am. Psychol.* **39**, 117–123 (1984).

25. Davidson, R. J. Prolegomenon to the structure of emotion: Gleanings from neuropsychology. *Cogn. Emot.* **6**, 245–268 (1992).

26. Levenson, R. W., Ekman, P. & Friesen, W. V. Voluntary Facial Action Generates Emotion-Specific Autonomic Nervous System Activity. *Psychophysiology* **27**, 363–384 (1990).

27. Johnston, V. S. The origin and function of pleasure. *Cognition and Emotion* **17**, 167–179 (2003).

28. Thorndike, E. L. *Animal intelligence; experimental studies,*. (The Macmillan Company, 1911). doi:10.5962/bhl.title.55072

29. Yin, H. H. & Knowlton, B. J. The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience* **7**, 464–476 (2006).

30. Goh, H. T., Gordon, J., Sullivan, K. J. & Winstein, C. J. Evaluation of attentional demands during motor learning: Validity of a dual-task probe paradigm. *J. Mot. Behav.* **46**, 95–105 (2014).

31. Lauwereyns, J. *et al.* Feature-based anticipation of cues that predict reward in monkey caudate nucleus. *Neuron* **33**, 463–73 (2002).

32. Lauwereyns, J., Watanabe, K., Coe, B. & Hikosaka, O. A neural correlate of response bias in monkey caudate nucleus. *Nature* **418**, 413–417 (2002).

33. Kawagoe, R., Takikawa, Y. & Hikosaka, O. Expectation of reward modulates cognitive signals in the basal ganglia. *Nat. Neurosci.* **1**, 411–416 (1998).

34. Pasupathy, A. & Miller, E. K. Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature* **433**, 873–876 (2005).

35. Yin, H. H. *et al.* Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. *Nat. Neurosci.* **12**, 333–341 (2009).

36. Thorn, C. A., Atallah, H., Howe, M. & Graybiel, A. M. Differential Dynamics of Activity Changes in Dorsolateral and Dorsomedial Striatal Loops during Learning. *Neuron* **66**,

781–795 (2010).

37. Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).

38. Smith, K. S. & Graybiel, A. M. Habit formation. *Dialogues Clin. Neurosci.* **18**, 33–43 (2016).

39. Bromberg-Martin, E. S., Matsumoto, M. & Hikosaka, O. Dopamine in Motivational Control: Rewarding, Aversive, and Alerting. *Neuron* **68**, 815–834 (2010).

40. Pawlak, V. & Kerr, J. N. D. Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. *J. Neurosci.* **28**, 2435–2446 (2008).

41. Goyal, N., Siddiqui, S., Chatterjee, U., Kumar, D. & Siddiqui, A. Neuropsychology of prefrontal cortex. *Indian J. Psychiatry* **50**, 202 (2008).

42. Eryilmaz, H. *et al.* Neural determinants of human goal-directed vs. habitual action control and their relation to trait motivation. *Sci. Rep.* **7**, (2017).

43. Patterson, T. K. & Knowlton, B. J. Subregional specificity in human striatal habit learning: a meta-analytic review of the fMRI literature. *Current Opinion in Behavioral Sciences* **20**, 75–82 (2018).

44. Graybiel, A. M. & Grafton, S. T. The striatum: Where skills and habits meet. *Cold Spring Harb. Perspect. Biol.* **7**, a021691 (2015).

45. Seger, C. A. & Spiering, B. J. A critical review of habit learning and the basal ganglia. *Front. Syst. Neurosci.* **5**, 1–9 (2011).

46. Aupperle, R. L., Melrose, A. J., Francisco, A., Paulus, M. P. & Stein, M. B. Neural substrates of approach-avoidance conflict decision-making. *Hum. Brain Mapp.* **36**, 449–62 (2015).

47. Aupperle, R. & Paulus, M. Neural systems underlying approach and avoidance in anxiety disorders. *Dialogues Clin. Neurosci.* **12**, 517–531 (2010).

48. Ernst, M., Pine, D. S. & Hardin, M. Triadic model of the neurobiology of motivated behavior in adolescence. *Psychological Medicine* **36**, 299–312 (2006).

49. Ernst, M. & Fudge, J. L. A developmental neurobiological model of motivated behavior: Anatomy, connectivity and ontogeny of the triadic nodes. *Neuroscience and Biobehavioral Reviews* **33**, 367–382 (2009).

50. Kozlik, J., Neumann, R. & Lozo, L. Contrasting motivational orientation and evaluative coding accounts: On the need to differentiate the effectors of approach/avoidance responses. *Frontiers in Psychology* **6**, 563 (2015).

51. Markman, A. B. & Brendl, C. M. Constraining Theories of Embodied Cognition. *Psychol. Sci.* **16**, 6–10 (2005).

52. Seibt, B., Neumann, R., Nussinson, R. & Strack, F. Movement direction or change in distance? Self- and object-related approach–avoidance motions. *J. Exp. Soc. Psychol.* **44**, 713–720 (2008).

53. Bamford, S. & Ward, R. Predispositions to Approach and Avoid Are Contextually Sensitive and Goal Dependent Susan. *Emotion* **8**, 174–183 (2008).

54. Saraiva, A. C., Schüür, F. & Bestmann, S. Emotional valence and contextual affordances flexibly shape approach-avoidance movements. *Front. Psychol.* **4**, 933 (2013).

55. De Houwer, J., Crombez, G., Baeyens, F. & Hermans, D. On the generality of the affective Simon effect. *Cogn. Emot.* **15**, 189–206 (2001).

56. Rinck, M. & Becker, E. S. Approach and avoidance in fear of spiders. *J. Behav. Ther. Exp. Psychiatry* **38**, 105–120 (2007).

57. Alexopoulos, T. & Ric, F. The evaluation-behavior link: Direct and beyond valence. *J. Exp. Soc. Psychol.* **43**, 1010–1016 (2007).

58. Huijding, J. *et al.* A behavioral route to dysfunctional representations: The effects of training approach or avoidance tendencies towards novel animals in children. *Behav. Res. Ther.* **47**, 471–477 (2009).

59. Huijding, J., Muris, P., Lester, K. J., Field, A. P. & Joosse, G. Training children to approach or avoid novel animals: Effects on self-reported attitudes and fear beliefs and information-seeking behaviors. *Behav. Res. Ther.* **49**, 606–613 (2011).

60. Woud, M. L., Becker, E. S., Lange, W.-G. & Rinck, M. Effects of Approach-Avoidance Training on Implicit and Explicit Evaluations of Neutral, Angry, and Smiling Face Stimuli. *Psychol. Rep.* **113**, 199–216 (2013).

61. Woud, M. L., Becker, E. S. & Rinck, M. Induction of implicit evaluation biases by approach- avoidance training: A commentary on Vandenbosch and De Houwer (this issue). *Cogn. Emot.* **25**, 1331–1338 (2011).

62. Becker, D., Jostmann, N. B., Wiers, R. W. & Holland, R. W. Approach avoidance training in the eating domain: Testing the effectiveness across three single session studies. *Appetite* **85**, 58–65 (2015).

63. Schakel, L. *et al.* The effects of a gamified approach avoidance training and verbal suggestions on food outcomes. *PLoS One* **13**, e0201309 (2018).

64. Warschburger, P., Gmeiner, M., Morawietz, M. & Rinck, M. Battle of plates: A pilot study of an approach-avoidance training for overweight children and adolescents. *Public Health Nutr.* **21**, 426–434 (2018).

65. Wittekind, C. E., Feist, A., Schneider, B. C., Moritz, S. & Fritzsche, A. The approach-avoidance task as an online intervention in cigarette smoking: A pilot study. *J. Behav. Ther. Exp. Psychiatry* **46**, 115–120 (2015).

66. Kakoschke, N., Kemps, E. & Tiggemann, M. The effect of combined avoidance and control training on implicit food evaluation and choice. *J. Behav. Ther. Exp. Psychiatry* **55**, 99–105 (2017).

67. Bijttebier, P., Beck, I., Claes, L. & Vandereycken, W. Gray's Reinforcement Sensitivity Theory as a framework for research on personality-psychopathology associations. *Clin. Psychol. Rev.* **29**, 421–430 (2009).

68. Heuer, K., Rinck, M. & Becker, E. S. Avoidance of emotional facial expressions in social anxiety: The Approach-Avoidance Task. *Behav. Res. Ther.* **45**, 2990–3001 (2007).

69. Peter J. Lang, Robert F. Simons, Marie Balaban, R. S. *Attention and Orienting: Sensory and Motivational Processes - Google Livros*. (1997).

70. Lange, W.-G., Keijsers, G., Becker, E. S. & Rinck, M. Social anxiety and evaluation of social crowds: explicit and implicit measures. *Behav. Res. Ther.* **46**, 932–43 (2008).

71. Roelofs, K. *et al.* Gaze direction differentially affects avoidance tendencies to happy and

angry faces in socially anxious individuals. *Behav. Res. Ther.* **48**, 290–294 (2010).

72. Horley, K., Williams, L. M., Gonsalvez, C. & Gordon, E. Face to face: visual scanpath evidence for abnormal processing of facial expressions in social phobia. *Psychiatry Res.* **127**, 43–53 (2004).

73. Horley, K., Williams, L. M., Gonsalvez, C. & Gordon, E. Social phobics do not see eye to eye:: A visual scanpath study of emotional expression processing. *J. Anxiety Disord.* **17**, 33–44 (2003).

74. Kuckertz, J. M., Strege, M. V & Amir, N. Intolerance for approach of ambiguity in social anxiety disorder. *Cogn. Emot.* **31**, 747–754 (2017).

75. Lange, W.-G., Allart, E., Keijsers, G. P. J., Rinck, M. & Becker, E. S. A Neutral Face Is Not Neutral Even if You Have Not Seen It: Social Anxiety Disorder and Affective Priming with Facial Expressions. *Cogn. Behav. Ther.* **41**, 108–118 (2012).

76. Struijs, S. Y. *et al.* Approach and avoidance tendencies in depression and anxiety disorders. *Psychiatry Res.* **256**, 475–481 (2017).

77. Bartoszek, G. & Winer, E. S. Spider-fearful individuals hesitantly approach threat, whereas depressed individuals do not persistently approach reward. *J. Behav. Ther. Exp. Psychiatry* **46**, 1–7 (2015).

78. Najmi, S., Kuckertz, J. M. & Amir, N. Automatic avoidance tendencies in individuals with contamination-related obsessive-compulsive symptoms. *Behav. Res. Ther.* **48**, 1058–1062 (2010).

79. Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S. & Lindenmeyer, J. Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychol. Sci.* **22**, 490–497 (2011).

80. Wiers, R. W., Rinck, M., Kordts, R., Houben, K. & Strack, F. Retraining automatic action-tendencies to approach alcohol in hazardous drinkers. *Addiction* **105**, 279–287 (2010).

81. Eberl, C. *et al.* Approach bias modification in alcohol dependence: Do clinical effects replicate and for whom does it work best? *Dev. Cogn. Neurosci.* **4**, 38–51 (2013).

82. Sharbanee, J. M. *et al.* The effect of approach/avoidance training on alcohol consumption is mediated by change in alcohol action tendency. *PLoS One* **9**, e85855 (2014).

83. Leeman, R. F. *et al. A Test of Multisession Automatic Action Tendency Retraining to Reduce Alcohol Consumption Among Young Adults in the Context of a Human Laboratory Paradigm*. *Alcoholism: Clinical and Experimental Research* **42**, (2018).

84. Goodman, W. K., Grice, D. E., Lapidus, K. A. B. & Coffey, B. J. Obsessive-compulsive disorder. *Psychiatr. Clin. North Am.* **37**, 257–267 (2014).

85. De Haan, S., Rietveld, E. & Denys, D. On the nature of obsessions and compulsions. *Mod. Trends Pharmacopsychiatry* **29**, 1–15 (2013).

86. Huppert, J. D. *et al.* The OCI-R: Validation of the subscales in a clinical sample. *J. Anxiety Disord.* **21**, 394–406 (2007).

87. Schwartzman, C. M. *et al.* Symptom subtype and quality of life in obsessive-compulsive disorder. *Psychiatry Res.* **249**, 307–310 (2017).

88. Abramowitz, J. S., Taylor, S. & McKay, D. Obsessive-compulsive disorder. *The Lancet*

374, 491–499 (2009).

89. Burguière, E., Monteiro, P., Mallet, L., Feng, G. & Graybiel, A. M. Striatal circuits, habits, and implications for obsessive-compulsive disorder. *Current Opinion in Neurobiology* **30**, 59–65 (2015).

90. Gillan, C. M. & Robbins, T. W. Goal-directed learning and obsessive–compulsive disorder. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, (2014).

91. Goodwin, G. M. The overlap between anxiety, depression, and obsessive-compulsive disorder. *Dialogues Clin. Neurosci.* **17**, 249–60 (2015).

92. Hartmann, A. & Millet, B. Repetitive movements and behaviors in neurological and psychiatric practice: Distinctions and similarities between Tourette disorder and obsessive–compulsive disorder. *Rev. Neurol. (Paris).* **174**, 199–202 (2018).

93. Kazhungil, F. & Mohandas, E. Management of obsessive-compulsive disorder comorbid with bipolar disorder. *Indian J. Psychiatry* **58**, 259 (2016).

94. Abramovitch, A., Dar, R., Mittelman, A. & Wilhelm, S. Comorbidity Between Attention Deficit/Hyperactivity Disorder and Obsessive-Compulsive Disorder Across the Lifespan. *Harv. Rev. Psychiatry* **23**, 245–262 (2015).

95. Abramowitz, J. S., Taylor, S. & McKay, D. Potentials and limitations of cognitive treatments for obsessive-compulsive disorder. *Cogn. Behav. Ther.* **34**, 140–147 (2005).

96. Lewin, A. B., Wu, M. S., McGuire, J. F. & Storch, E. A. Cognitive behavior therapy for obsessive-compulsive and related disorders. *Psychiatr. Clin. North Am.* **37**, 415–445 (2014).

97. Rowa, K., Antony, M. M. & Swinson, R. P. Exposure and Response Prevention. in *Psychological treatment of obsessive-compulsive disorder: Fundamentals and beyond.* 79–109 (American Psychological Association, 2007). doi:10.1037/11543-004

98. McKay, D. *et al.* Efficacy of CBT for obsessive-compulsive disorder. *Psychiatry Res.* **227**, 104–113 (2015).

99. Kozak, M. J. & Foa, E. B. Obsessions, overvalued ideas, and delusions in obsessive-compulsive disorder. *Behav. Res. Ther.* **32**, 343–353 (1994).

100. Veale, D. Over-valued ideas: a conceptual analysis. *Behav. Res. Ther.* **40**, 383–400 (2002).

101. Abramowitz, J. S. Variants of exposure and response prevention in the treatment of obsessive-compulsive disorder: A meta-analysis. *Behav. Ther.* **27**, 583–600 (1996).

102. Christensen, H., Hadzi-Pavlovic, D., Andrews, G. & Mattick, R. Behavior therapy and tricyclic medication in the treatment of obsessive-compulsive disorder: A quantitative review. *J. Consult. Clin. Psychol.* **55**, 701–711 (1987).

103. Rosa-Alcázar, A. I., Sánchez-Meca, J., Gómez-Conesa, A. & Marín-Martínez, F. Psychological treatment of obsessive-compulsive disorder: A meta-analysis. *Clinical Psychology Review* **28**, 1310–1325 (2008).

104. Olatunji, B. O., Davis, M. L., Powers, M. B. & Smits, J. A. J. Cognitive-behavioral therapy for obsessive-compulsive disorder: A meta-analysis of treatment outcome and moderators. *J. Psychiatr. Res.* **47**, 33–41 (2013).

105. Soomro, G. M., Altman, D. G., Rajagopal, S. & Oakley Browne, M. Selective serotonin re-

uptake inhibitors (SSRIs) versus placebo for obsessive compulsive disorder (OCD). *Cochrane Database Syst. Rev.* CD001765 (2008). doi:10.1002/14651858.CD001765.pub3

106. Pittenger, C. & Bloch, M. H. Pharmacological treatment of obsessive-compulsive disorder. *Psychiatric Clinics of North America* **37**, 375–391 (2014).

107. Stewart, S. E. *et al.* Meta-analysis of association between obsessive-compulsive disorder and the 3′ region of neuronal glutamate transporter gene *SLC1A1*. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **162**, 367–379 (2013).

108. Brennan, B. P., Rauch, S. L., Jensen, J. E., Pope, H. G. & Jr. A critical review of magnetic resonance spectroscopy studies of obsessive-compulsive disorder. *Biol. Psychiatry* **73**, 24–31 (2013).

109. Bhattacharyya, S. *et al.* Anti-Brain Autoantibodies and Altered Excitatory Neurotransmitters in Obsessive–Compulsive Disorder. *Neuropsychopharmacology* **34**, 2489–2496 (2009).

110. Chakrabarty, K., Bhattacharyya, S., Christopher, R. & Khanna, S. Glutamatergic Dysfunction in OCD. *Neuropsychopharmacology* **30**, 1735–1740 (2005).

111. Pittenger, C., Bloch, M. H. & Williams, K. Glutamate abnormalities in obsessive compulsive disorder: Neurobiology, pathophysiology, and treatment. *Pharmacology and Therapeutics* **132**, 314–332 (2011).

112. Stewart, S. E. *et al.* A Single-Blinded Case-Control Study of Memantine in Severe Obsessive-Compulsive Disorder. *J. Clin. Psychopharmacol.* **30**, 34–39 (2010).

113. Greenberg, W. M. *et al.* Adjunctive glycine in the treatment of obsessive-compulsive disorder in adults. *J. Psychiatr. Res.* **43**, 664–670 (2009).

114. Myers, K. M., Carlezon, W. A. & Davis, M. Glutamate receptors in extinction and extinction-based therapies for psychiatric illness. *Neuropsychopharmacology* **36**, 274–293 (2011).

115. Norberg, M. M., Krystal, J. H. & Tolin, D. F. A Meta-Analysis of D-Cycloserine and the Facilitation of Fear Extinction and Exposure Therapy. *Biol. Psychiatry* **63**, 1118–1126 (2008).

116. Foa, E. B. *et al.* Randomized, Placebo-Controlled Trial of Exposure and Ritual Prevention, Clomipramine, and Their Combination in the Treatment of Obsessive-Compulsive Disorder. *Am. J. Psychiatry* **162**, 151–161 (2005).

117. Quality of life for patients with obsessive-compulsive disorder. *Am. J. Psychiatry* **153**, 783–788 (1996).

118. Markarian, Y. *et al.* Multiple pathways to functional impairment in obsessive–compulsive disorder. *Clin. Psychol. Rev.* **30**, 78–88 (2010).

119. Schruers, K., Koning, K., Luermans, J., Haack, M. J. & Griez, E. Obsessive-compulsive disorder: A critical review of therapeutic perspectives. *Acta Psychiatrica Scandinavica* **111**, 261–271 (2005).

120. Gillan, C. M., Robbins, T. W., Sahakian, B. J., van den Heuvel, O. A. & van Wingen, G. The role of habit in compulsivity. *Eur. Neuropsychopharmacol.* **26**, 828–840 (2016).

121. Figee, M. *et al.* Compulsivity in obsessive-compulsive disorder and addictions. *Eur.*

Neuropsychopharmacol. **26**, 856–868 (2016).

122. Chambers, C. D., Garavan, H. & Bellgrove, M. A. Insights into the neural basis of response inhibition from cognitive and clinical neuroscience. *Neurosci. Biobehav. Rev.* **33**, 631–646 (2009).

123. Struijs, S. Y. *et al.* The predictive value of Approach and Avoidance tendencies on the onset and course of depression and anxiety disorders. *Depress. Anxiety* **35**, 551–559 (2018).

124. Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S. & Lindenmeyer, J. Retraining Automatic Action Tendencies Changes Alcoholic Patients' Approach Bias for Alcohol and Improves Treatment Outcome. *Psychol. Sci.* **22**, 490–497 (2011).

125. Taylor, C. T. & Amir, N. Modifying automatic approach action tendencies in individuals with elevated social anxiety symptoms. *Behav. Res. Ther.* **50**, 529–536 (2012).

126. Asnaani, A., Rinck, M., Becker, E. & Hofmann, S. G. The Effects of Approach–Avoidance Modification on Social Anxiety Disorder: A Pilot Study. *Cognit. Ther. Res.* **38**, 226–238 (2014).

127. Amir, N., Kuckertz, J. M. & Najmi, S. The effect of modifying automatic action tendencies on overt avoidance behaviors. *Emotion* **13**, 478–84 (2013).

128. Weil, R., Feist, A., Moritz, S. & Wittekind, C. E. Approaching contamination-related stimuli with an implicit Approach-Avoidance Task: Can it reduce OCD symptoms? An online pilot study. *J. Behav. Ther. Exp. Psychiatry* **57**, 180–188 (2017).

129. Henriques, T. Overcoming automatic response tendencies : behavioral findings and computational model-based analysis Thesis to obtain the Master of Science Degree in Biomedical Engineering. (Instituto Superior Técnico, 2015).

130. Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–42 (1997).

131. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–6 (1997).

132. August, T. The R User Conference , useR ! 2011 August 16-18 2011 University of Warwick , Coventry , UK Book of Contributed Abstracts Contents. *Stanford Encycl. Philos.* (2011).

133. Lin, T. Y. *et al.* Microsoft COCO: Common objects in context. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **8693 LNCS**, 740–755 (2014).

134. Bertron, A. *et al.* International Affective Picture System (IAPS): Technical Manual and Affective Ratings Lang , P . J ., Bradley , M . M ., & Cuthbert , B . N . NIMH Center for the Study of Emotion and Attention 1997 with the assistance over the years of . - Mark Greenwal. *Int. Affect. Pict. Syst.* (1997).

135. Smarr, K. L. & Keefer, A. L. Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionna. *Arthritis Care Res. (Hoboken).* **63**, S454–S466 (2011).

136. Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H. & Covi, L. The Hopkins Symptom Checklist (HSCL): a self-report symptom inventory. *Behav. Sci.* **19**, 1–15 (1974).

137. Carver, C. S. & White, T. L. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *J. Pers. Soc. Psychol.* **67**, 319–333 (1994).

138. McCrae, R. R. & Costa, P. T. A contemplated revision of the NEO Five-Factor Inventory. *Pers. Individ. Dif.* **36**, 587–596 (2004).

139. Rassin, E. The White Bear Suppression Inventory (WBSI) Focuses on Failing Suppression Attempts. *Eur. J. Pers.* **17**, 285–298 (2003).

140. John, R. & Julie, D. The Positive and Negative Affect Schedule (PANAS): Construct validity , measurement properties and normative data in a large non-clinical sample. *Br. J. Clin. Psychol.* **43**, 245–65 (2004).

141. Harkness, J. a & Schoua-Glusberg, A. Questionnaires in Translation. *ZUMA-Nachrichten Spez.* 87–126 (1998).

142. European Social Survey. ESS Round 7 Translation Guidelines. 49 (2014).

143. Canavarro, M. C. S. *Inventário de Sintomas Psicopatológicos - B.S.I.* (Faculdade de Psicologia e de Ciências da Educação da UC, 1999).

144. Derogatis, L. R. & Melisaratos, N. The Brief Symptom Inventory: an introductory report. *Psychol. Med.* **13**, 595 (1983).

145. Souza, F. P., Foa, E. B., Meyer, E., Niederauer, K. G. & Cordioli, A. V. Psychometric properties of the Brazilian Portuguese version of the Obsessive-Compulsive Inventory – Revised (OCI-R). *Rev. Bras. Psiquiatr.* **33**, 137–143 (2011).

146. Cardoso, I. Propriedades Psicométricas da Versão Portuguesa do Obsessive – Compulsive Inventory — Revised Orientador : Professor Doutor Miguel Faria Universidade Lusófona de Humanidades e Tecnologias Escola de Psicologia e Ciências da. *Psicologia Clínica* (Universidade Lusófona de Humanidades e Tecnologias, 2015).

147. Cardoso, I. Propriedades psicométricas da versão portuguesa do obsessive–compulsive inventory—revised. (2016). Available at: http://recil.grupolusofona.pt/handle/10437/6816?show=full. (Accessed: 13th September 2019)

148. Faria, M. & Cardoso, I. Propriedades psicométricas da versão portuguesa do Obsessive-Compulsive Inventory – Revised. **1**, 91–100 (2017).

149. Campos, R. C. & Gonçalves, B. The portuguese version of the beck depression inventory-II (BDI-II) preliminary psychometric data with two nonclinical samples. *Eur. J. Psychol. Assess.* **27**, 258–264 (2011).

150. Magalhães, E. *et al.* NEO-FFI : Psychometric Properties of a Short Personality Inventory in Portuguese Context. **27**, 599–614 (2005).

151. Galinha, I., Pereira, C. R. & Esteves, F. Versão reduzida da escala de afeto positivo e negativo portuguesa - PANAS-Port-VR: Análise fatorial confirmatória e invariância temporal. *Psicologia* **28**, 53–65 (2014).

152. Fabião, C., Barbosa, A., Fleming, M. & Silva, M. C. Instrumentos de rastreio de somatização em geral e perturbações somatoformes: Nos cuidados primários. *Acta Medica Portuguesa* **24**, 439–448 (2011).

153. Foa, E. B. *et al.* The Obsessive-Complusive Inventory: Development and validation of a

short version. *Psychol. Assess.* **14**, 485–495 (2002).

154. Whelan, R. Effective Analysis of Reaction Time Data. *Psychol. Rec.* **58**, 475–482 (2008).

155. Miller, J. Short Report: Reaction Time Analysis with Outlier Exclusion: Bias Varies with Sample Size. *Q. J. Exp. Psychol. Sect. A* **43**, 907–912 (1991).

156. Singmann, H., Spieler, D. H. & Schumacher, E. Running head : MIXED MODELS 1 An Introduction to Mixed Models for Experimental Psychology. (2017).

157. Lindgren, K. P. *et al.* Attempted training of alcohol approach and drinking identity associations in us undergraduate drinkers: Null results from two studies. *PLoS One* **10**, e0134642 (2015).

158. Wheaton, M. G., Gershkovich, M., Gallagher, T., Foa, E. B. & Simpson, H. B. Behavioral avoidance predicts treatment outcome with exposure and response prevention for obsessive–compulsive disorder. *Depress. Anxiety* **35**, 256–263 (2018).

159. Gillan, C. M. *et al.* Enhanced avoidance habits in obsessive-compulsive disorder. *Biol. Psychiatry* **75**, 631–8 (2014).

160. Starcevic, V. *et al.* The Nature and Correlates of Avoidance in Obsessive–Compulsive Disorder. *Aust. New Zeal. J. Psychiatry* **45**, 871–879 (2011).

161. Manos, R. C. *et al.* The impact of experiential avoidance and obsessive beliefs on obsessive-compulsive symptoms in a severe clinical sample. *J. Anxiety Disord.* **24**, 700–708 (2010).

162. Eberl, C. *et al.* Implementation of Approach Bias Re-Training in Alcoholism-How Many Sessions are Needed? *Alcohol. Clin. Exp. Res.* **38**, 587–594 (2014).

163. Dajani, D. R. & Uddin, L. Q. Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. *Trends in Neurosciences* **38**, 571–578 (2015).

164. Mackie, M. A., Van Dam, N. T. & Fan, J. Cognitive control and attentional functions. *Brain Cogn.* **82**, 301–312 (2013).

165. Gruner, P. & Pittenger, C. Cognitive inflexibility in Obsessive-Compulsive Disorder. *Neuroscience* **345**, 243–255 (2017).

166. Francazio, S. K. & Flessner, C. A. Cognitive flexibility differentiates young adults exhibiting obsessive-compulsive behaviors from controls. *Psychiatry Res.* **228**, 185–190 (2015).

167. Sternheim, L., Van Der Burgh, M., Berkhout, L. J., Dekker, M. R. & Ruiter, C. Health and Disability Poor cognitive flexibility, and the experience thereof, in a subclinical sample of female students with obsessive-compulsive symptoms. (2014). doi:10.1111/sjop.12163

168. Sheridan, M. R. Response programming, response production, and fractionated reaction time. *Psychol. Res.* **46**, 33–47 (1984).

169. Keele, S. W. Behavioral Analysis of Movement. in *Comprehensive Physiology* (John Wiley & Sons, Inc., 1981). doi:10.1002/cphy.cp010231

170. Gentier, I. *et al.* A comparative study of performance in simple and choice reaction time tasks between obese and healthy-weight children. *Res. Dev. Disabil.* **34**, 2635–2641 (2013).

171. Towey, J. *et al.* Endogenous event-related potentials in obsessive-compulsive disorder. *Biol. Psychiatry* **28**, 92–98 (1990).

172. Miyata, A. *et al.* Event-related potentials in patients with obsessive-compulsive disorder. *Psychiatry Clin. Neurosci.* **52**, 513–518 (1998).

173. Yamamuro, K. *et al.* A longitudinal, event-related potential pilot study of adult obsessive-compulsive disorder with 1-year follow-up. *Neuropsychiatr. Dis. Treat.* **12**, 2463–2471 (2016).

174. Mehl, N., Mueller-Wieland, L., Mathar, D. & Horstmann, A. Retraining automatic action tendencies in obesity. *Physiol. Behav.* **192**, 50–58 (2018).

175. Jalal, B. *et al.* Novel Smartphone Interventions Improve Cognitive Flexibility and Obsessive-Compulsive Disorder Symptoms in Individuals with Contamination Fears. *Sci. Rep.* **8**, 14923 (2018).

176. Sharbanee, J. M. *et al.* The effect of approach/avoidance training on alcohol consumption is mediated by change in alcohol action tendency. *PLoS One* **9**, e85855 (2014).

177. Ernst, L. H., Weidner, A., Ehlis, A.-C. & Fallgatter, A. J. Controlled attention allocation mediates the relation between goal-oriented pursuit and approach–avoidance reactions to negative stimuli. *Biol. Psychol.* **91**, 312–320 (2012).

178. Rapp, A. M., Bergman, R. L., Piacentini, J. & Mcguire, J. F. Evidence-Based Assessment of Obsessive–Compulsive Disorder. *J. Cent. Nerv. Syst. Dis.* **8**, JCNSD.S38359 (2016).

179. Castro-Rodrigues, P. *et al.* Criterion Validity of the Yale-Brown Obsessive-Compulsive Scale Second Edition for Diagnosis of Obsessive-Compulsive Disorder in Adults. *Front. Psychiatry* **9**, (2018).

180. Goodman, W. K. *et al.* The Yale-Brown Obsessive Compulsive Scale: I. Development, Use, and Reliability. *Arch. Gen. Psychiatry* **46**, 1006–1011 (1989).

181. Kessler, R. C. *et al.* Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry* **6**, 168–76 (2007).

182. Pinto-Gouveia, J., & Albuquerque, P. (2007). Versão Portuguesa do Inventário de Supressão do Urso Branco. Unpublished manuscprit.

6.Annexes

Content:

- ✓ ***Estimation of the Number of Participants for Screening***
- ✓ ***Exploratory Analysis with Data from Previous Study***
- ✓ ***Pilot Study: AAT Stimuli Selection***

6.1 Estimation of the Number of Participants for Screening

6.1.1 Introduction

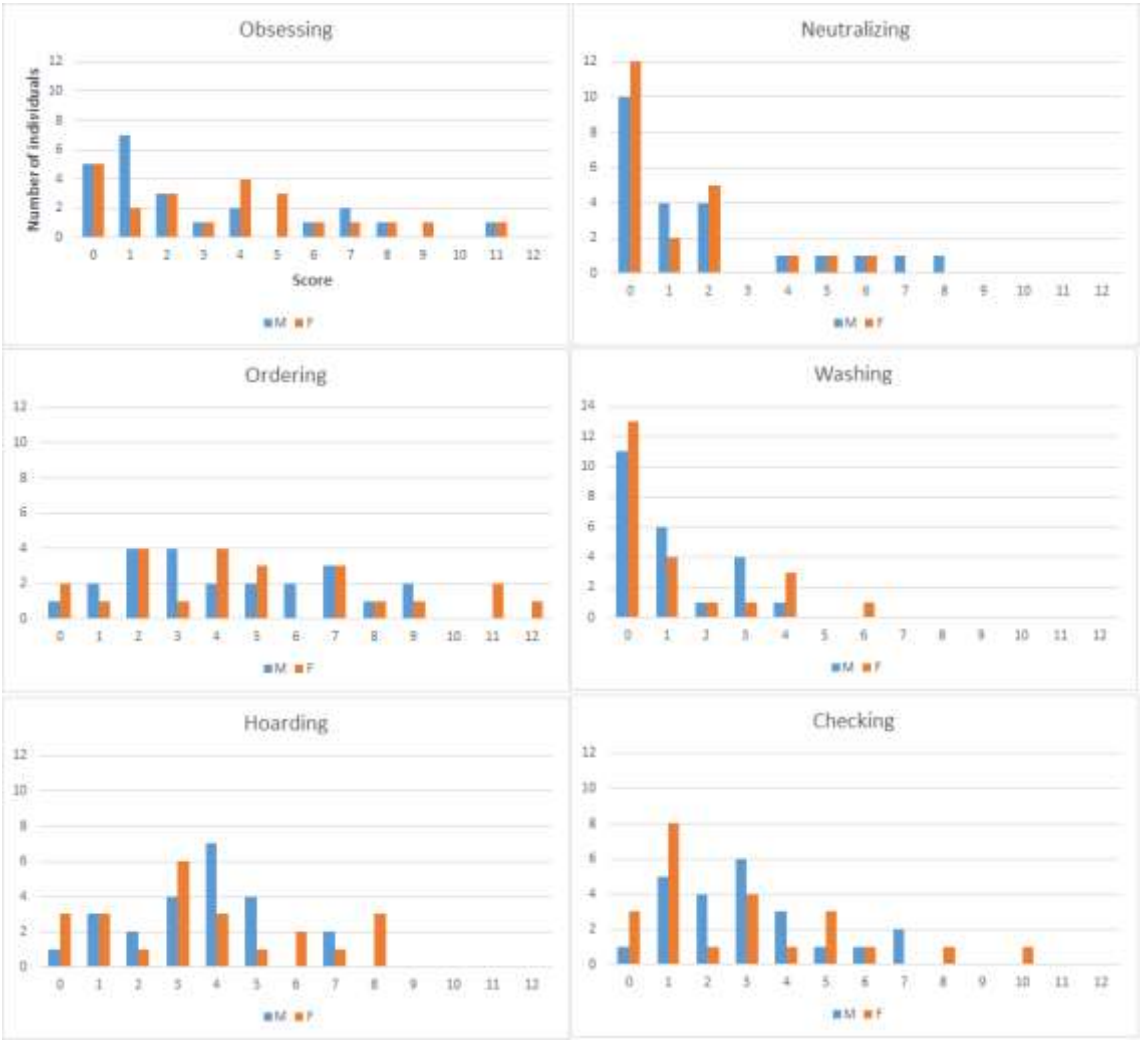
The main objective established at the beginning of the current thesis was to behaviourally train two groups of healthy participants with low versus high fear of contamination traits, in order to gradually build-up new approach responses to contamination-related stimuli. Thus, an estimation of the number of individuals with higher versus low fear of contamination traits in an average student population was performed, in order to know beforehand the general number of students needed to screen. This was done by consulting the data of a previous study in the lab[129] which also applied the OCI-R (explained in detail in section 2.3 Sample Description) to a group of healthy students. In addition, it is also important to note that the data obtained via the OCI-R in this previous study had not been analysed at the time.

6.1.2 Methods

For each participant, all OCI-R subscale scores were calculated (*Checking* - Items 2, 8 and 14; *Hoarding* - 1, 7, 13; *Washing* - 5, 11, 17; *Ordering* - 3, 9, 15; *Mental Neutralizing* - 4, 10, 16; *Obsessing* - 6, 12, 18)[86,153]. Afterwards, the number of participants displaying each score (ranging from 0 to 12) was calculated for each of the six OCI-R subscales. Taking into account the evidence suggesting that the OCI-R scores might vary with gender[153], the frequency of scores was split into Male and Female.

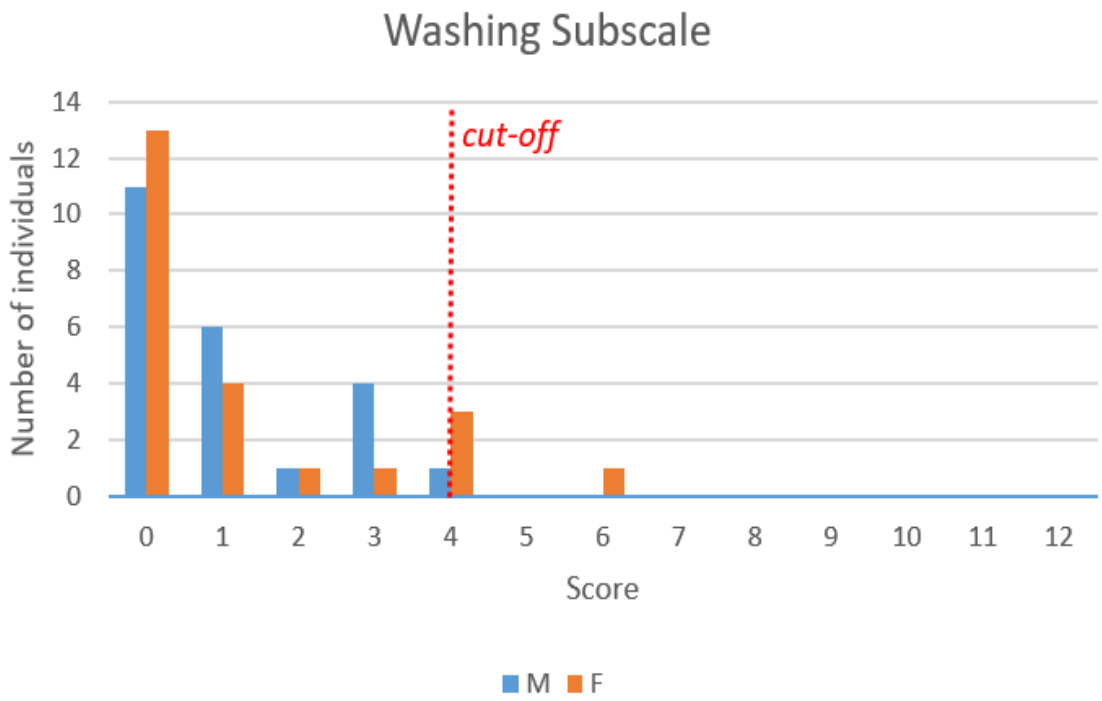
6.1.3 Results and Conclusions

The sample tested in this previous study displayed the lowest average score for the *Washing* scale ($\overline{X}$ = 1.11, ∂ = 1.20), compared to all other subscales (*Mental Neutralizing*: $\overline{X}$ = 1.5, ∂ = 1.63; *Checking*: $\overline{X}$ = 2.89, ∂ = 1.77; *Hoarding*: $\overline{X}$ = 3.57, ∂ = 1.74; *Ordering*: $\overline{X}$ = 4.65, ∂ = 2.53; *Obsessing*: $\overline{X}$ = 3.17, ∂ = 2.55).



Annexed Figure 1: Frequency of Scores in each OCI-R Subscale - Number of subjects (y axis) for each possible score, ranging from 0 to 12 (x axis), for each of the six subscales of the Obsessive Compulsive Inventory-Revised (OCI-R). The tested sample was part of a previous study where participants, aged between 18-29 years old, were asked to fill in several questionnaires, including the OCI-R. All participants were healthy students. M - Male (blue bars); F - Female (orange bars).

Moreover, visual inspection of the figure below indicated that only 5 out of 47 participants, approximately one tenth of participants, displayed a Washing Subscale score above the cut-off that was described in the literature to adequately separate participants with low vs high fear of contamination[86,153] (see section *2.3 Sample Description* in the Methods for a more detailed explanation about the cut-off). Since the intention was to collect 25 participants with a Washing Subscale score above the cut-off, this indicated that screening more or less 250 participants would allow to achieve that number of candidates. Therefore, after this analysis was performed, it was decided to use an online questionnaire in order to facilitate the screening and pre-selection of participants for the AAT training.



Annexed Figure 2: Frequency of Scores in the Washing Subscale - Number of subjects (y axis) for each possible score, ranging from 0 to 12 (x axis) in the washing subscale in the OCI-R questionnaire. The cut-off score established for this subscale was 4 (red dotted vertical line).

6.2 Exploratory Analyses with Data from a Previous Study

6.2.1 Image Ratings Analyses at the First Assessment

6.2.1.1 Introduction

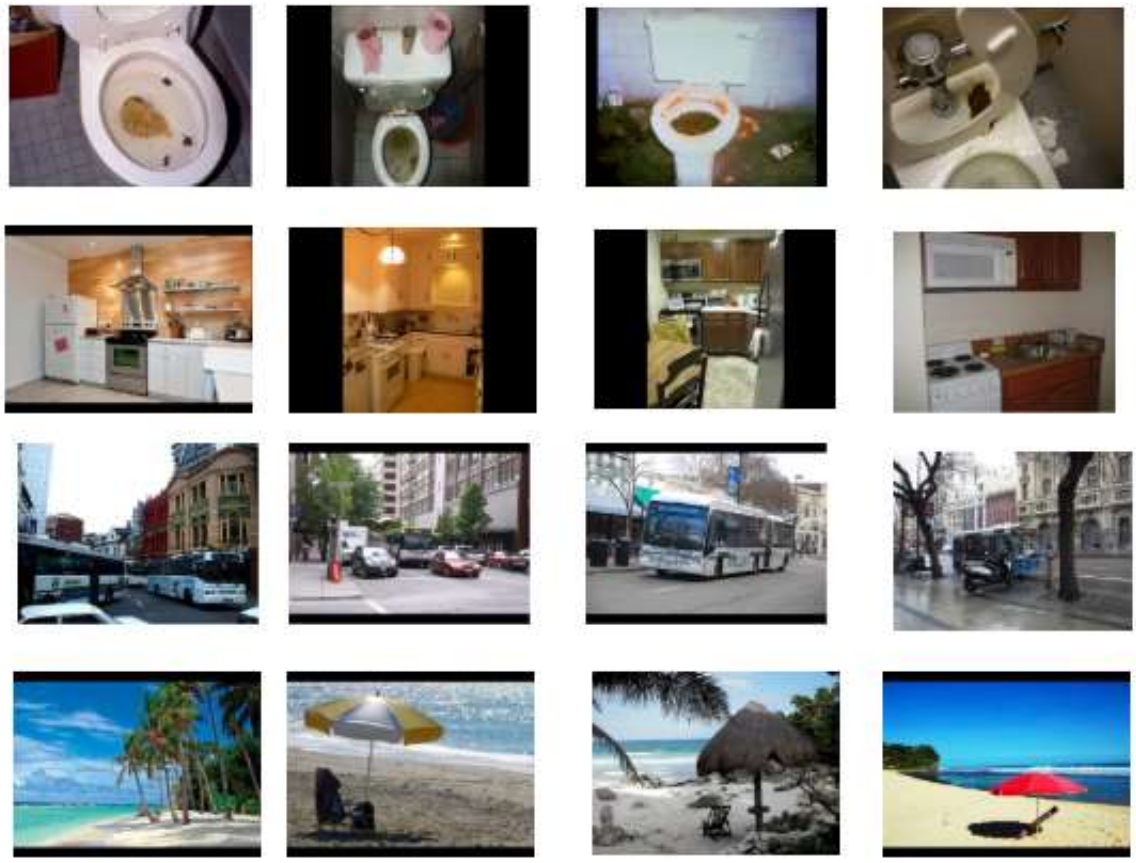
Results of a previous AAT Training study performed in the lab indicated that the stimuli used there needed improvement[129]. Thus, to select new appropriate stimuli to be used in the current study, an additional analysis of these data was performed to examine the average ratings (Pleasant/Unpleasant) by all participants for each image upon their initial examination, i.e., before the first training session, in order to assess the images in term of their adequacy and quality. This brief analysis focused on two (negative and neutral-kitchen) of the original four categories, since, for the current thesis, only two (*approach negative* and *avoid neutral-kitchen*) of the original four training conditions were of interest.

In addition, correlations between the average ratings for the negative images (before training) *versus* the OCI-R and Washing Subscale scores were performed, in order to assess how the obsessive-compulsive traits were related to the reported degree of unpleasantness of the negative images. The expectation here was that the more

strongly participants had reported to have OC-like and fear of contamination traits, the more unpleasant they would rate the contamination-related images.

6.2.1.2 Methods

The image ratings' data at the 1st assessment, as well as the two aforementioned correlations were performed and analysed with the open source software R Studio.



Annexed Figure 3: Images used in the Previous Study - The first row displays the contamination-related images (number 1 to 4) used for the negative stimuli. The second row displays the kitchen images (numbers 9 to 12) used in the neutral-kitchen stimuli. The third row displays the street-related images (numbers 5 to 8) used in the neutral-street stimuli. The fourth row displays the beach-related images (numbers 13 to 16) used in the positive stimuli.

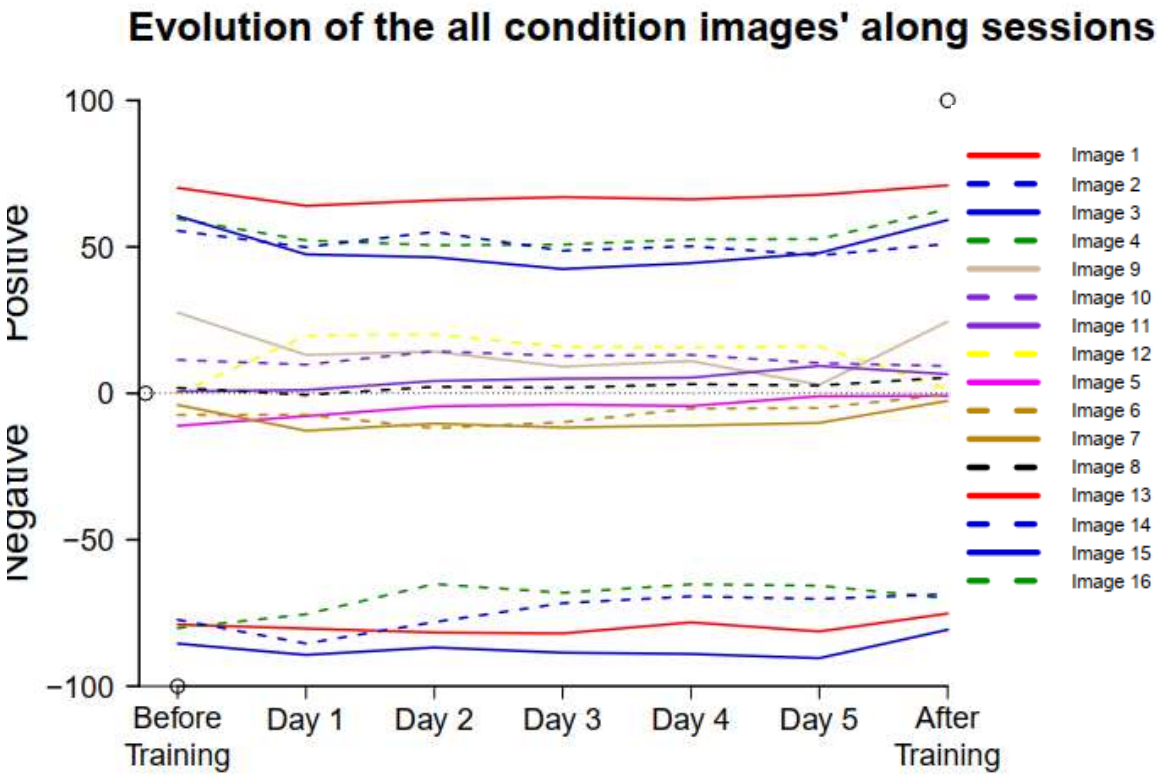
6.2.1.4 Results and Conclusions

Figure 4 below depicts a more detailed version of the "*Evolution of the ratings along sessions*" graphic used in the previous study (see [129], page 62), that was made in order to check and assess possible differences between images, that were not observable in the summarized graphic in the former thesis. Thus, visual inspection of figure 4 was combined with a discussion of the characteristics of the used images with regard to three features: (1) Easy to recognize (2) Strength of content and (3) Typicality of content:

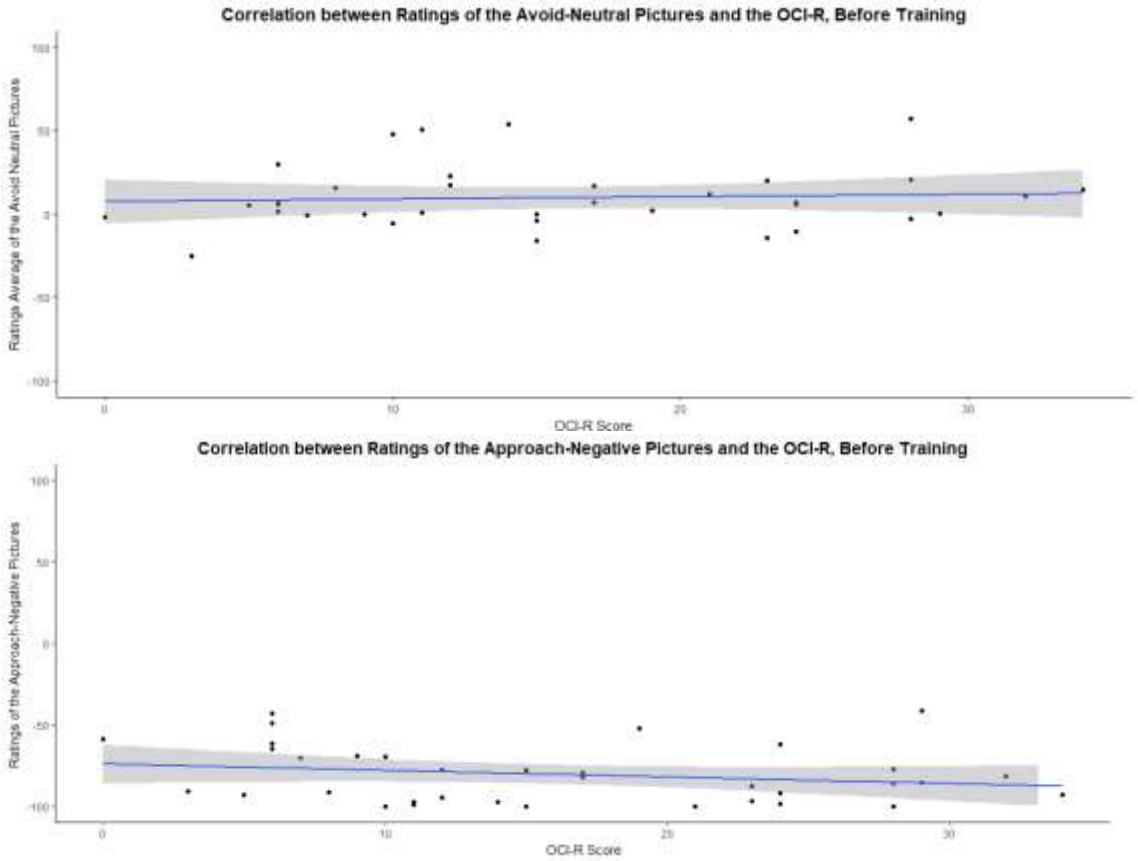
Negative Category - Image 15 seemed to have been rated as more negative than the other negative pictures, which was in line with our subjective impression that the

surroundings of the depicted toilet elicited strong associations to a drug-related environment, with the presence of empty bottles and an unwashed, filthy floor. Thus, for the improved negative images of the current thesis, the aim established was to find images that still depicted dirty toilets and evoked fear of contamination, but that did not emphasize any other negative aspects than just moderate dirty surroundings.

Neutral-Kitchen Category - Image 9 seemed to have been rated higher than the others. Visual inspection revealed that it depicted a very modern and sophisticated kitchen equipment. Thus, for the improved neutral-kitchen images the aim established was to find neutral-kitchen images that depicted common equipment without any distraction objects such as food or plants.

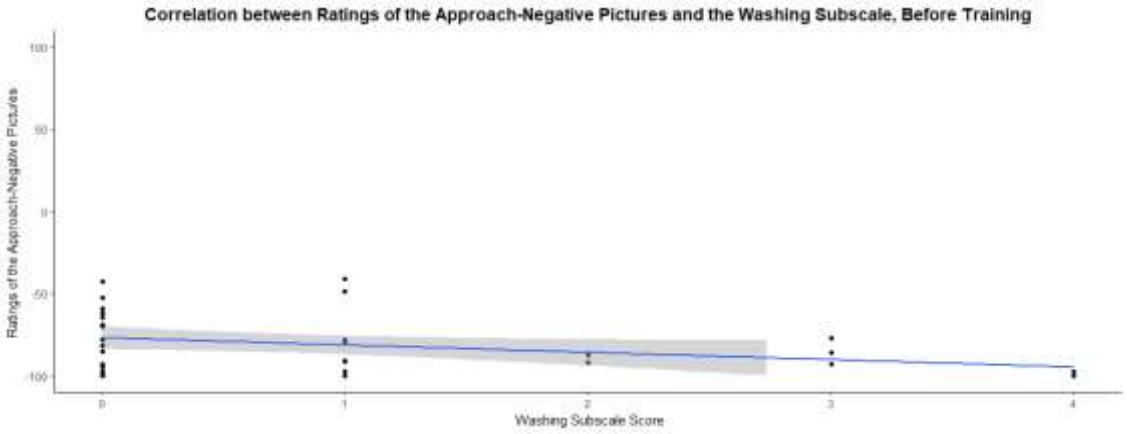


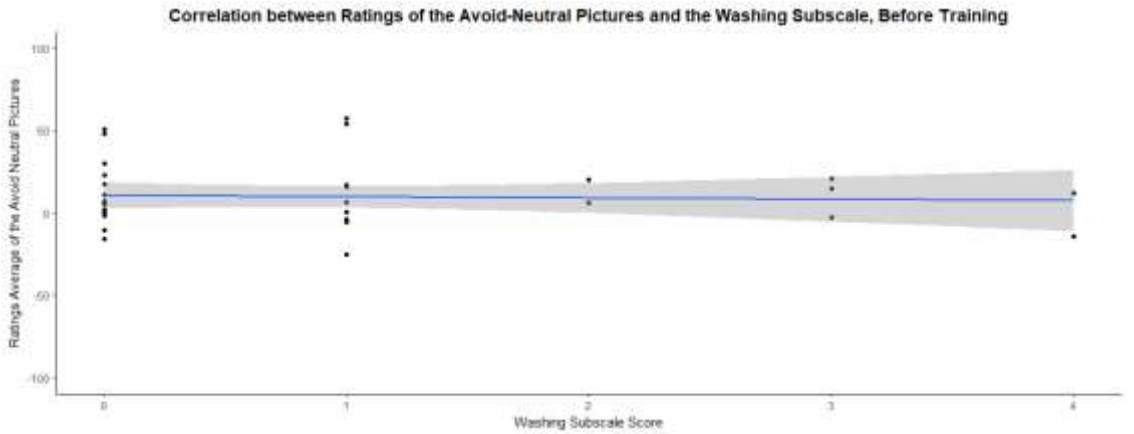
Annexed Figure 4: Image Ratings from All Participants Along the Approach-Avoidance Task Sessions - In each AAT session, participants were asked to rate the pictures between -100 to 100. Each line represents the average ratings the respective image across the AAT protocol. Approach-Negative Images: Image 1 to Image 4; Avoid-Neutral Images: Image 9 to Image 12; Approach-Neutral Images: Image 5 to Image 8; Avoid-Positive Images: Image 13 to Image 16.



Annexed Figure 5: Correlation between the OCI-R scores, Ratings for the Negative Images and for the Neutral Images before Training - Each black dot represents one participant, with his/her respective average rating for the images (y axis) and the corresponding OCI-R score of the participant (x axis). In the upper panel, results showed a statistically non-significant correlation (r = -0.21, p = 0.21; confidence intervals = 95%). In the lower panel, results also showed a statistically non-significant correlation (r = 0.069, p = 0.69; confidence intervals = 95%). According to these results, together with visual inspection of these correlations, indicated that OCI-R individuals score differences did not influence the ratings in the negative and neutral images.

Regarding the correlations above, results showed that independently of the OCI-R score the majority of participants rated the negative images as being unpleasant, i.e., close to *"-100*" mark (r = -0.21, p = 0.21; confidence intervals = 95%), and the neutral as being neither unpleasant nor pleasant, i.e., close to 0 (r = 0.069, p = 0.69; confidence intervals = 95%).





Annexed Figure 6: Correlation between the Washing Subscale Scores, Ratings for the Negative Images and for the Neutral Images before Training - Each black dot represents one participant, with his/her respective average rating for the images (y axis) and the corresponding Washing score of the participant (x axis). In the upper panel, results showed a trend (r = -0.31, p = 0.066; confidence intervals = 95%). In the lower panel, results also showed a statistically non-significant correlation (r = 0.049, p = 0.78; confidence intervals = 95%). According to these results, participants with higher washing score tended to rate the negative as more unpleasant, while rating the neutral images close to zero.

Regarding the correlation between the Washing Subscale and the negative images, the first correlation above showed a weak tendency that pointed in the direction to the expectation (r = -0.31, p-value = 0.066; confidence intervals = 95%), indicating that participants whose Washing Subscale scores were higher tended to rate the negative images as being more unpleasant, compared to participants whose washing subscale scores were lower. As for the neutral-kitchen images, the second correlation above indicated that the majority of participants rated the neutral-kitchen images close to zero, independent of their washing subscale score, as expected (r = 0.049, p = 0.78; confidence intervals = 95%).

In short, the two sets of correlations are in agreement with each other: Regarding the negative images, these results show that there was a weak tendency that people with high Washing and OCI-R scores tended to experience negative images as being more unpleasant, due to their higher contamination-related traits, compared to participants with lower OCI-R and Washing scores (OCI-R papers). Moreover, the non-significant correlation with the OCI-R could be explained by the fact that the OCI-R questionnaire measures all symptoms associated with all six subtypes of OCD, hence its correlation might add information from other subtypes that are not directly related to automatic reactions to contamination-related stimuli. On the contrary, the correlation with the Washing Subscale showed a trend, which is in line with the fact that it measures specifically the fear of contamination traits. On the other hand, the results show that the avoid neutral pictures were rated as being neutral, which confirms their initial labelling as *neutral*.

6.2.2 Reaction Times and Reaction Biases before Training

6.2.2.1 Introduction

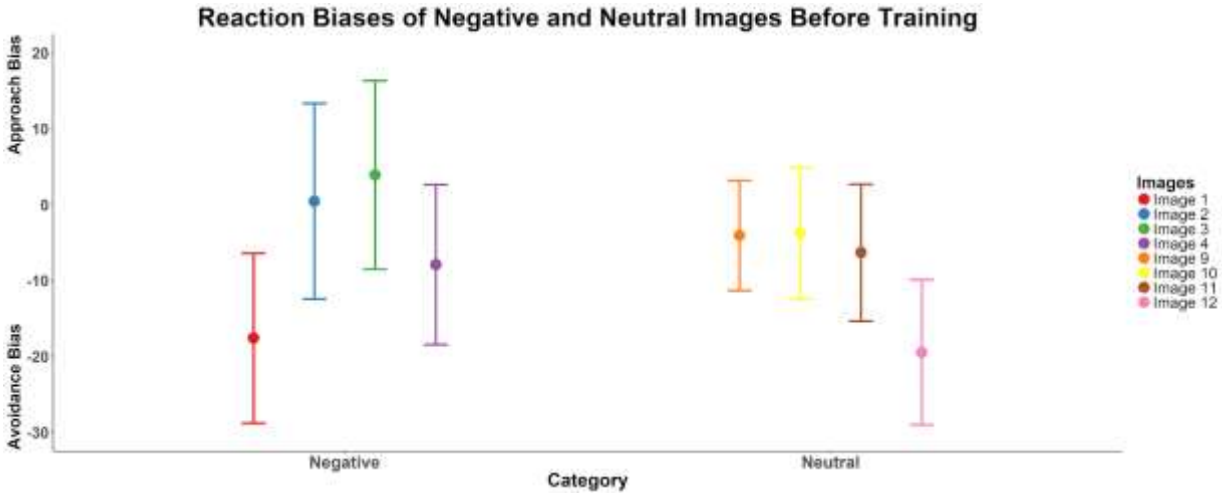
As part of the additional analysis performed on the data of the previous study[129], in order to select new appropriate stimuli for the current thesis, the RTs and RBs that all participants displayed for the negative and neutral-kitchen images, before training, were briefly analysed. In the same fashion as in the ratings analysis, correlations between the RBs *versus* the OCI-R and Washing Subscales score were performed, as well as a correlation between the RBs the ratings.

For the first set of analyses, the expectation was that the negative images would have elicited an avoidance RB due to their contamination-related content, whereas the neutral-kitchen images would display neither an approach nor an avoidance RB due to their relatively neutral content. As for the aforementioned correlations, it was expected that the higher the OCI-R and Washing scores, the slower participants would be when approaching negative images compared to avoid, and, consequently, the higher their RBs.

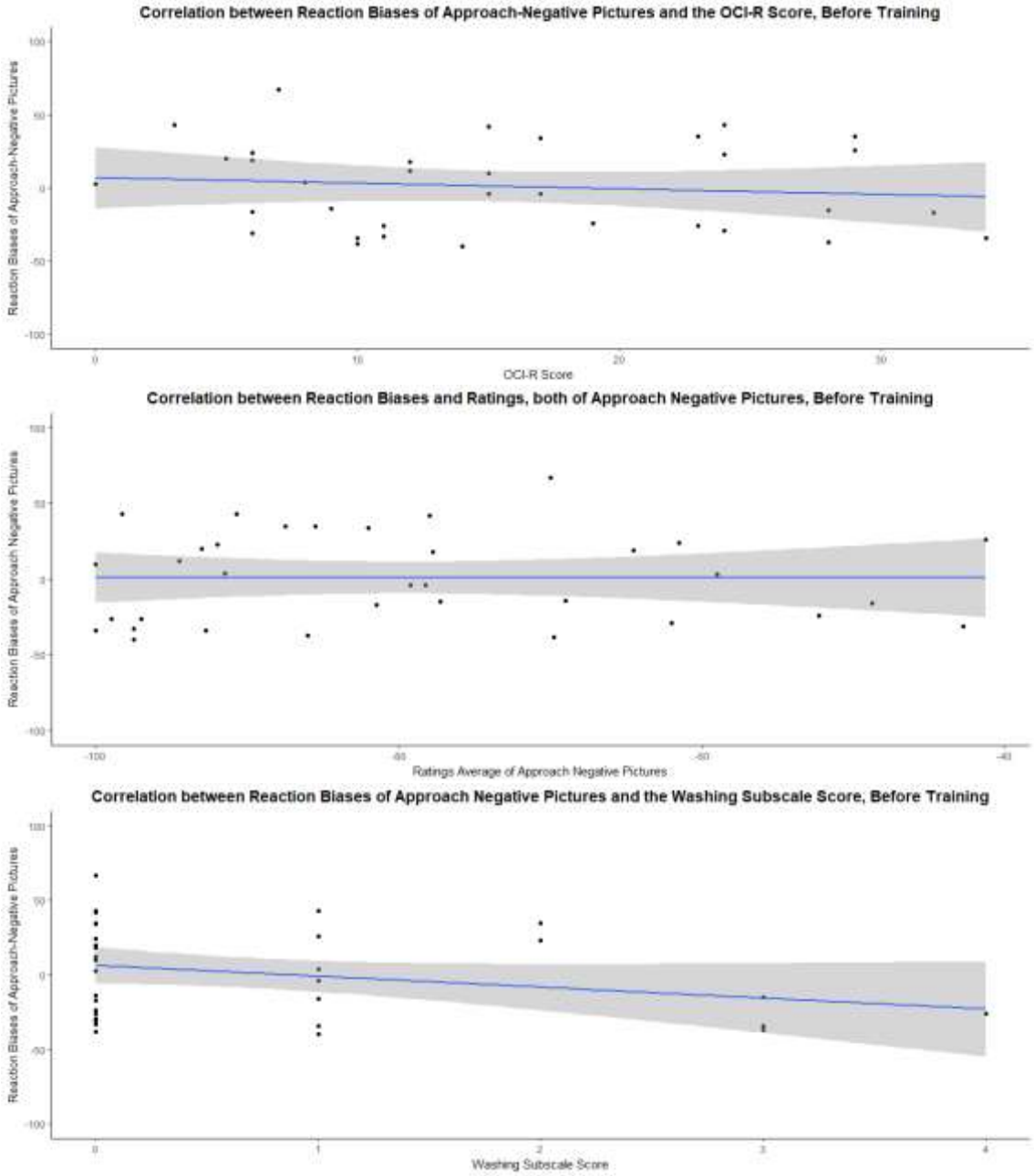
6.2.2.2 Methods

The same methodology used in the ratings analyses was used here.

6.2.2.3 Results and Conclusions



Annexed Figure 7: Average Reaction Biases Displayed for each Negative and Neutral Image before Training - The images 1 to 4 and 9 to 12 represent the negative and neutral images, respectively. Visual inspection of the figure suggests that the negative images, overall, did not elicit any reaction biases, except for the image 1, who elicited a relatively small avoidance reaction bias. Similarly, the neutral images did not elicit any reaction biases, except for the picture 12, which elicited a relatively small avoidance reaction bias.



Annexed Figure 8: Correlation between the Reaction Biases Displayed for the Negative Images, Washing Subscale Scores, OCI-R Scores and Ratings for the Negative Images before Training - In the first correlation results showed a weak but statistically non-significant correlation (r = -0.24, p = 0.17; confidence intervals = 95%). Despite of the missing statistical significance, the result of this correlation was cautiously interpreted for exploratory reasons: According to this correlation, the higher participants' OCI-R score, the stronger the avoidance reaction biases tended to be. In the second correlation results showed that he higher participants' Washing score, the stronger the avoidance reaction biases participants displayed for the negative images (r = -0.42, p = 0.011; confidence interval = 95%). In the third correlation, results showed no significant correlation between the reaction biases and the ratings for the negative images (r = 0.19, p = 0.28; confidence interval = 95%).

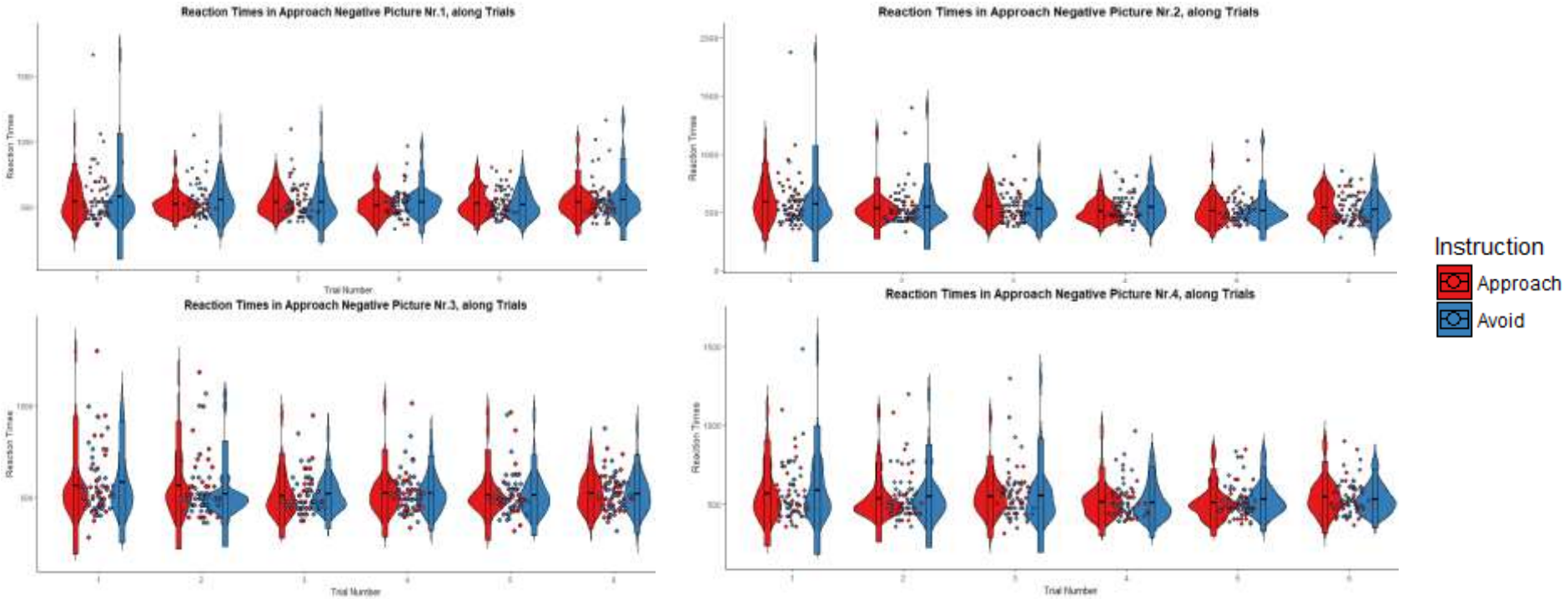
Regarding the RBs displayed for the negative images, in contrast to the hypothesis, visual inspection of figure 7 (see above) indicated that, overall, these images did not elicit a strong avoidance RB, except for a relatively weak avoidance RB for image

number 1. These results could be explained by an enhanced cognitive attention control when participants were repeatedly shown contamination-related images, throughout the 1st assessment session. As for the RBs displayed for the neutral-kitchen images, in line with the expectations, these images did not seem to elicit any specific strong reaction, except for image 8 that seemed to have elicited a small avoidance RB.

Regarding the correlations between the OCI-R and the RBs displayed for the negative images, pointing towards the expectation, results indicated that the higher the OCI-R score, the stronger the avoidance RBs for these images ($r = -0.24$, $p = 0.17$, confidence intervals = 95%; Upper panel in figure 8). As argued in the correlation with the ratings above, this non-statistical significance might be explained by the fact that the OCI-R questionnaire measures all symptoms associated with all six subtypes of OCD, as opposed to specifically the fear of contamination subtype. Indeed, results showed a significant moderate correlation between the Washing Subscale scores and the RBs displayed for the negative images ($r = -0.42$, $p = 0.011$, confidence interval = 95%; Middle panel in figure 8), indicating that the higher the Washing Subscale score, the stronger the avoidance RB participants displayed for the negative images. As for the correlation between the RBs and the ratings for the negative images, results showed a weak non-significant correlation in the direction of the expectation, indicating a tendency that the higher the avoidance RBs displayed ($r = 0.1$, $p = 0.28$, confidence interval = 95%; Lower panel in figure 8) the more unpleasant participants seemed to rate the negative images.

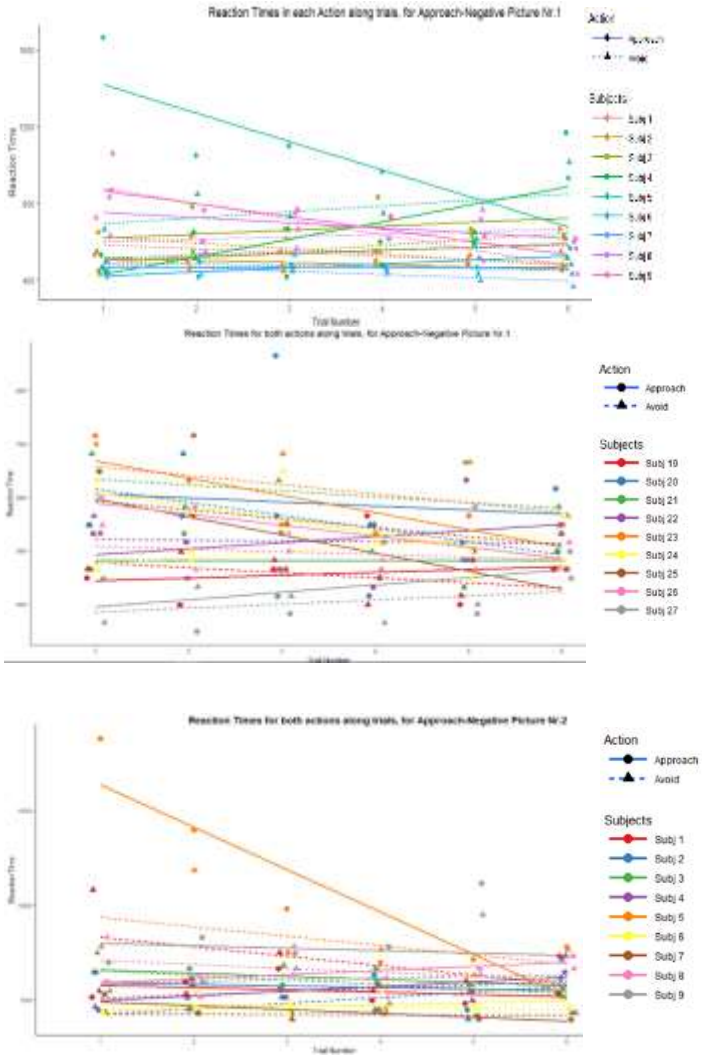
Thus, when briefly looking the results of the analysis performed here, both the OCI-R and washing scores seem to be related, although weakly, to the RBs and ratings participants displayed for the negative images. This could be attributed to the fact that the sample used in this study was not pre-selected according to any specific OCD-related criteria, and, therefore, any correlation with an OCD-related measurement could have been halted. As such, this indicated the importance of using inclusion criteria in the current thesis, such as with the washing subscale of the OCI-R, in a way that would assess OCD-like traits, particularly to pre-select participants according to low and high fear of contamination traits. Moreover, this would allow to better test how training OCD-related stimuli would affect participants with OCD-like traits.

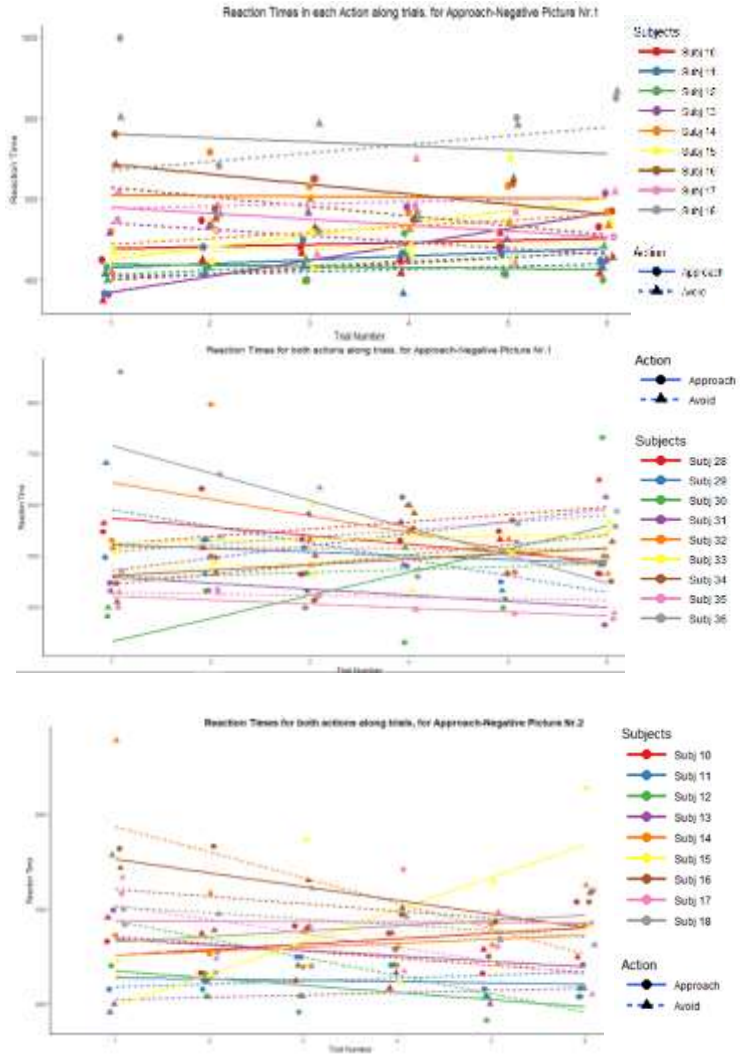
As an additional step of the RTs analyses, the average RTs when avoiding *versus* approaching negative images were analysed throughout trials at the 1st assessment, in order to assess any possible differences. To do so, violin-plots were performed to visualize the distribution of the RTs for each response and compare it between trials. The expectation was that the average RTs while approaching negative images would be higher than when avoiding them, since incongruent responses (approaching a negative stimuli) requires additional mental resources compared to congruent responses.

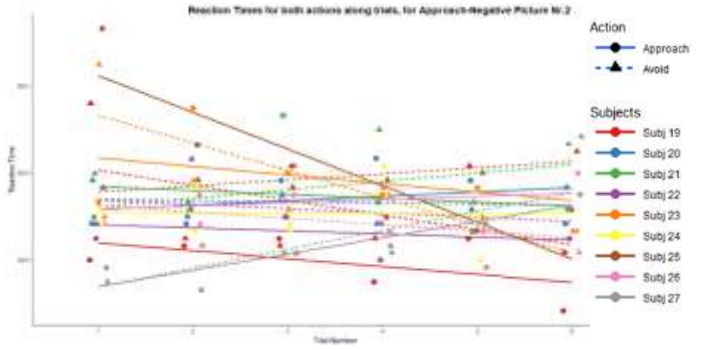


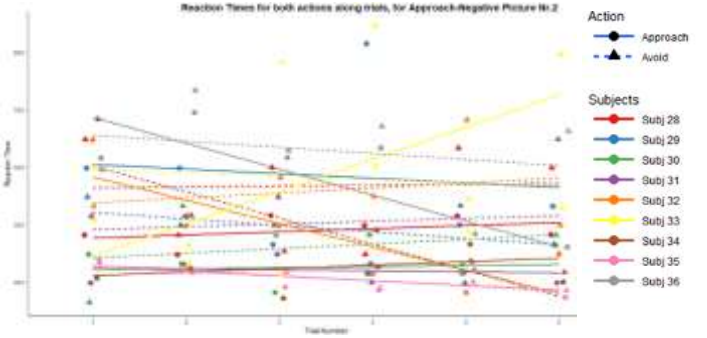
Annexed Figure 9: Viol*in Plots of the Reaction Times a*long *Trials in each Negative Image b*efore Training - At the 1st assessment session, participants had to approach and avoid the same images 6 times each, whereby the trials for an image were randomized among the other images' trials. The width of a violin plot describes the distribution density of the variable of interest (reaction times): wider sections of the violin plots represent a higher probability that participants would have a given reaction times value, while thinner sections represent the opposite. The vertical straight bars describe the standard deviations, while the black lines inside represent the means, both concerning the reaction times values. The dots represent the individual reaction times, i.e., the time it took each participants to perform in a given instruction. Blue - Avoid; Red - Approach.

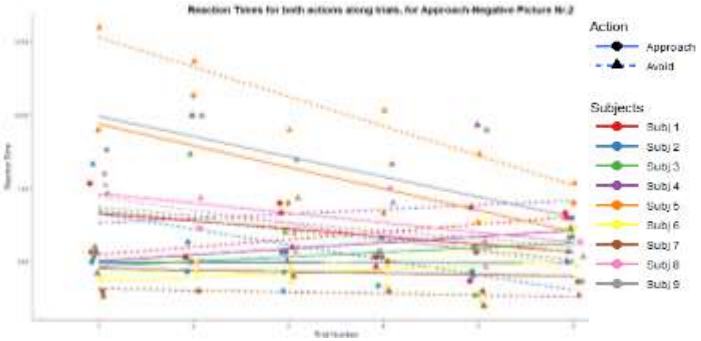
In contrast to the expectations, visual inspection of the violin plots above seemed to indicate that the RTs when approaching *versus* avoiding the negative images were not different throughout the trials, at the first assessment, for any of the negative images. To further analyse these RTs differences between approach and avoidance, a similar analysis was performed to visualize the progression of the RTs but now at the individual level.

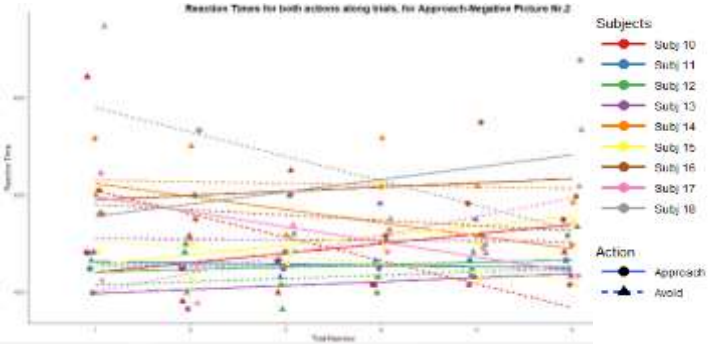

Reaction Times in each Action along trials, for Approach-Negative Picture Nr.1
Reaction Times for both actions along trials, for Approach-Negative Picture Nr.1
Reaction Times for both actions along trials, for Approach-Negative Picture Nr.2
Action
Approach
Avoid
Subjects
Subj 1
Subj 2
Subj 3
Subj 4
Subj 5
Subj 6
Subj 7
Subj 8
Subj 9
Subj 10
Subj 11
Subj 12
Subj 13
Subj 14
Subj 15
Subj 16
Subj 17
Subj 18
Subj 19
Subj 20
Subj 21
Subj 22
Subj 23
Subj 24
Subj 25
Subj 26
Subj 27

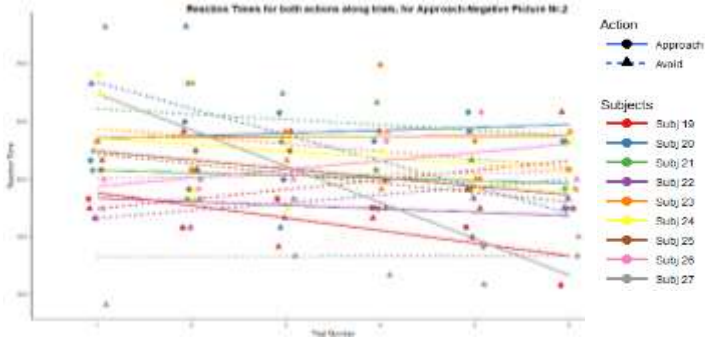

Reaction Times in each Action along trials, for Approach-Negative Picture Nr.1
Reaction Times for both actions along trials, for Approach-Negative Picture Nr.1
Reaction Times for both actions along trials, for Approach-Negative Picture Nr.2
Action
Approach
Avoid
Subjects
Subj 10
Subj 11
Subj 12
Subj 13
Subj 14
Subj 15
Subj 16
Subj 17
Subj 18
Subj 28
Subj 29
Subj 30
Subj 31
Subj 32
Subj 33
Subj 34
Subj 35
Subj 36

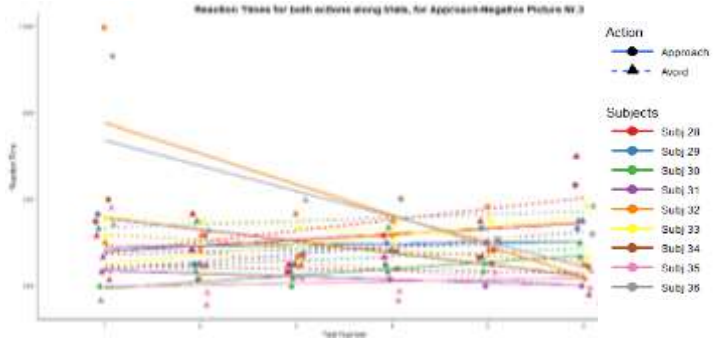

Action
Approach
Avoid
Subjects
Subj 19
Subj 20
Subj 21
Subj 22
Subj 23
Subj 24
Subj 25
Subj 26
Subj 27

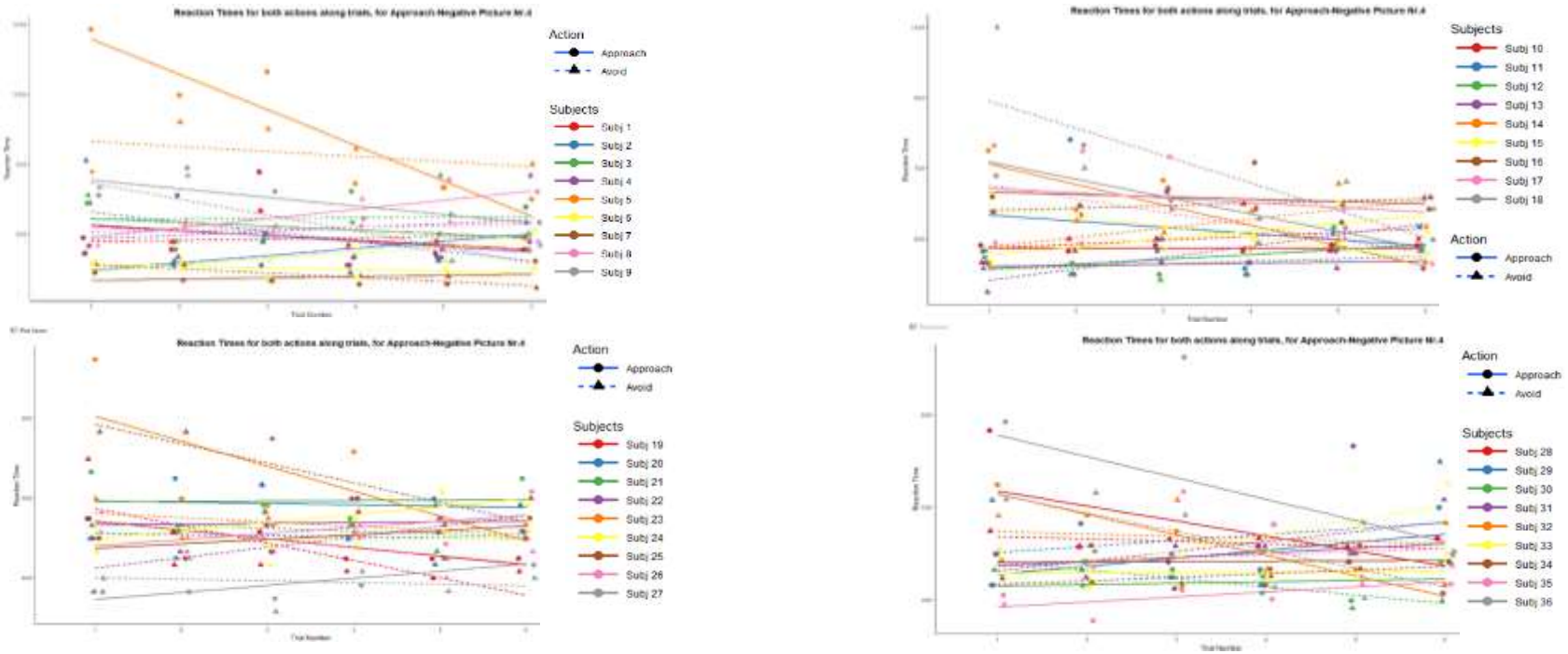

Action
Approach
Avoid
Subjects
Subj 28
Subj 29
Subj 30
Subj 31
Subj 32
Subj 33
Subj 34
Subj 35
Subj 36


Action
Approach
Avoid
Subjects
Subj 1
Subj 2
Subj 3
Subj 4
Subj 5
Subj 6
Subj 7
Subj 8
Subj 9


Subjects
Subj 10
Subj 11
Subj 12
Subj 13
Subj 14
Subj 15
Subj 16
Subj 17
Subj 18
Action
Approach
Avoid


Action
Approach
Avoid
Subjects
Subj 19
Subj 20
Subj 21
Subj 22
Subj 23
Subj 24
Subj 25
Subj 26
Subj 27


Action
Approach
Avoid
Subjects
Subj 28
Subj 29
Subj 30
Subj 31
Subj 32
Subj 33
Subj 34
Subj 35
Subj 36



Annexed Figure 10*: Reaction Times a*long *Trials for Each Negative Image b*efore Training - Each colour depicts a different subject. The shape of the dots and lines represent an instruction: Both the dashed and the continuous line, along with the triangles and circles describe each participant's avoiding and approaching reaction times, respectively, throughout the 6 trials in each instruction. Each four sets of graphs represent the reaction times along the trials for specific images (first set - negative image number 1; second set - negative image number 2; third set - negative image number 3; fourth set - negative image number 4). Visual inspection of these results seems to indicate no common general pattern regarding the participants' initial approach and avoid tendencies and their evolution over time, for each negative image, before training.

In contrast to the expectations, visual inspection of the figures above does not indicate a common pattern regarding participants' approach and avoidance tendencies throughout the trials of the 1st assessment. In fact, although some participants seemed to display a tendency to react faster when avoiding compared to when approaching the negative images, throughout the trials, as expected, others seemed to have different patterns: (1) a tendency to react faster when approaching compared to avoiding negative images; (2) a tendency to approach faster initially but then become faster at avoiding negative images; (3) a tendency to react equally fast when approach or avoiding the negative images.

Thus, visual inspection of the RTs distribution for the approach versus avoidance responses at the group and individual level led to the suggestion the negative stimuli might have elicited a generally enhanced attention level, thereby preventing differences between approach and avoid negative. Moreover, it was considered that preceding images could have influenced the RTs for the negative images, or vice-versa. As such, it was decided to restrict the way the trials were automatically shuffled for each participant in the current thesis, so that: (1) There could not be two consecutive negative trials; and (2) That the trials preceding the trials negative pictures had to follow a specific pattern of repetitions that kept changing across the total 192 trials (see section *2.2.2 Summary of Modification* in the Methods for a more detailed explanation).

6.3 Pilot Study: AAT Stimuli Selection

6.3.1 Introduction

The long-term aim of the project that includes the current thesis is to apply an AAT Training Protocol as an add-on therapy to ExRPT in OCD patients. As such, the first step of the current thesis consisted of choosing appropriate stimuli to be used for the AAT, since a previous study in the lab reported that the stimuli used in that work needed improvement[129]**.** Those results were briefly re-analysed and fine-tuned in this thesis (see section *6.2 Exploratory Analysis with Data from Previous Study* in the Annexes). Thus, in order to improve the quality of the stimuli, it was decided to search for new stimuli that fulfilled the following criteria: (1) easy to recognize; (2) strong content; (3) typical, i.e., whose content could be found in an everyday environment. With these characteristics, it was expected that the newly added stimuli would more reliably elicit automatic tendencies during the AAT, in a controlled laboratory environment.

To find new appropriate images that fulfilled the criteria mentioned above, a web-based search using the Common Object in Context (COCO) database[133] was performed, after which the selected pictures were assessed via an online questionnaire that was filled in by a sample of students who did not participate in any of the other parts of this thesis. Participants rated each image on a scale from 1 to 10 with regard to three questions: (1) Comprehensibility (Easy/Hard); (2) Strength of content

(Pleasant/Unpleasant); and (3) Reaction elicited (Approach/Avoid). The purpose of these online ratings was to briefly analyse participants' evaluations and select the images that had the most appropriate features. Thus, the focus was set on choosing negative images for the AAT Training, since this study aimed at investigating the effects of regular behavioral training (AAT Training) on participants' reactions when not paying directly attention to the content of the images (AAT Assessment). More precisely, it was hypothesized that the more salient the features of the negative images were in the training, the stronger the responses displayed would be and the less likely any misinterpretation would be.

At the time the online ratings were performed, the plan was to implement only two of the original four training conditions in this thesis (*approach negative* and *avoid neutral-kitchen*), hence both the web-based search and online ratings concerned only two categories of images, negative and neutral-kitchen images. The latter was assessed to make sure their content were perceived as neutral.

With regard to the framework applied, it was expected for both categories to be rated as being easy to recognize (close to the mark "1" of the ratings). Specifically for the negative images, it was anticipated that they would to be rated as being unpleasant and eliciting a strong avoidance response (both close to the mark "10" of the ratings). In an opposite pattern, it was expected for the neutral-kitchen images to be rated as neither pleasant nor unpleasant, and eliciting a reaction in the middle of the "*approach*" and "*avoid*" ends, (both around the "5" mark of the ratings).

6.3.2 Methods

6.3.2.1 Stimuli

The Common Object in COntext (COCO) database, which allows image-search with specific objects due to its object detection and segmentation algorithms, was used to search for new images to be used as stimuli for the AAT, specifically images displaying dirty toilets for the negative category, and kitchen-related scenarios for the neutral kitchen category. For the negative category, images exhibiting the following characteristics were excluded: Absence of a toilet; Sideways toilet or at the periphery of the image; Drug-related environments; Extreme dirty, unwashed or unusual surroundings. For the neutral-kitchen category, images displaying the following characteristics were excluded: Positive content such as food and colorful walls; Intense sunlight or artificial light; Small environment. As a result of the COCO-based search, together with some of the previously used images of a previous study[129], 8 negative and 8 neutral-kitchen images were selected.

Afterwards, all images were edited using the software *Photoshop* so that they all had equal dimensions (400x300 pixels) and did not display content that could distract participants from the centre of the images, such as dirty, unwashed floor, plants, foods or red-coloured objects.

According to the features explained above, visual inspection of the images retained allowed to pre-select the best images of each category: 6 out of the 8 negative images and 4 out of the 8 neutral-kitchen images. These pre-selected images were afterwards used for the online ratings.

6.3.2.2 Online Questionnaire

An email stating the main objective of the current thesis and the link to access the web page of the online ratings was sent to the Communication Office of University of Algarve, who in turn sent it to the respective faculties. The link redirected the users to a webpage where they had to click on one of possible four links, which corresponded to four different versions of the questionnaire. The versions only differed in the order of presentation of images within each category.

The structure of the entire online ratings was the following: (1) An introduction message, where it stated the main objective of the study, the instructions for the questionnaire itself and data privacy; Agree on the following sentence: "*I do not have a diagnostic or history of psychiatric or neurological disorders, as well as any type of chronic disorders*"; (2) An informed consent, that participants had to read and decide whether or not to agree in participating in the study; (3) Image ratings, starting with the neutral-kitchen and followed by the negative pictures; (4) Portuguese versions of the Obsessive Compulsive Inventory-Revised (OCI-R) and Brief Symptom Inventory (BSI) scales; (5) Final question whereupon participants were asked if they were interested in getting feedback regarding their BSI score.

Annexed Figure 11: Image Used in the Online Questionnaire - The images in the first row displays the negative (contamination-related images), while the second row displays the neutral (kitchen-related) images.

Annexed Figure 12: Images Ratings in the Online Questionnaire - In this case, a neutral-kitchen image is shown, whereupon participants had to rate the question "Quão fácil/difícil é perceber o conteúdo da imagens?" ("How easy/hard is it to understand the content of the image?") via a discrete scale from 1 ("Easy") to 10 ("Hard").

Participants had to rate each image with regard to three questions: "*Quão fácil/difícil é perceber o conteúdo da imagem?*" ("*How easy/hard it is to understand the image?*"), "*Quão agradável/desagradável é esta imagem para ti, neste momento?*" ("*How pleasant/unpleasant is this image for you, right now?*"), "*A tua reação a esta imagem é...*" ("*Your reaction to this image is…?*"). Participants would have to rate each image via a discrete 1 (Easy/Pleasant/Approach) to 10 (Hard/Unpleasant/Avoid) scale. Additionally, a fourth question would ask participants to describe each image in simple words.

6.3.2.3 Questionnaires: OCI-R and BSI

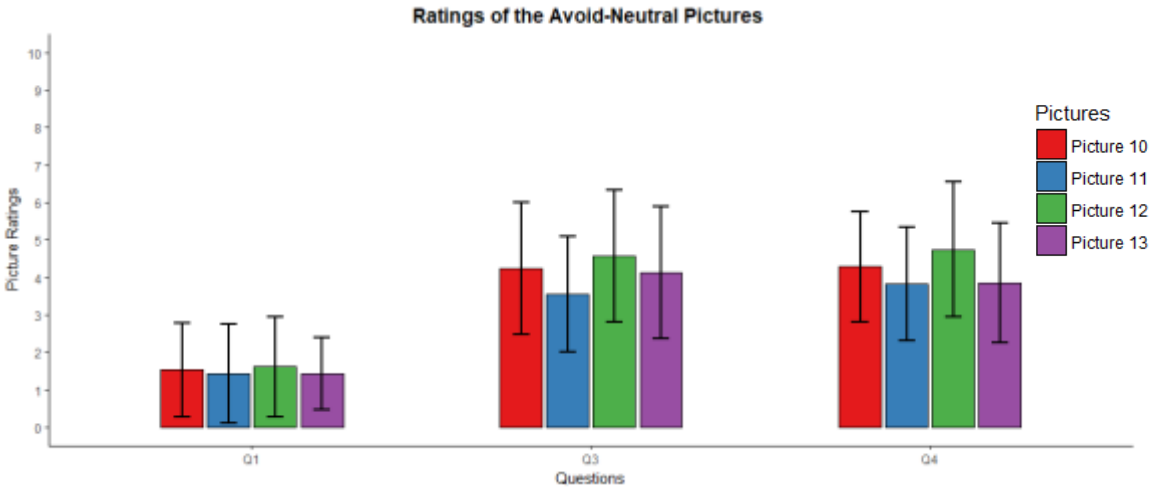
Besides the ratings, the Obsessive Compulsive Inventory-Revised (OCI-R) and Brief Symptom Inventory (BSI) were used in order to allow to analysing how the ratings depend on each participants' fear of contamination traits. The description of these two questionnaires is provided in the main text of the current thesis, in section *2.3 Sample Description* in the Methods.

Since the BSI provided a general but important status on the participants' mental well-being, it was decided to ask participants if they wanted to receive feedback regarding their BSI score via email. This email contained the scores of each symptom dimension, the BSI total scores, the respective *cut-off*, as well as an interpretation of the results, whereby we would summarize their level of mental well-being as "*Reduced"* or "*Without Significant Problems"*. For the participants who scored above the overall BSI cut-off, recommended entities were added in the email with whom they were advised to contact to further assess their mental well-being, in a more thorough way. No

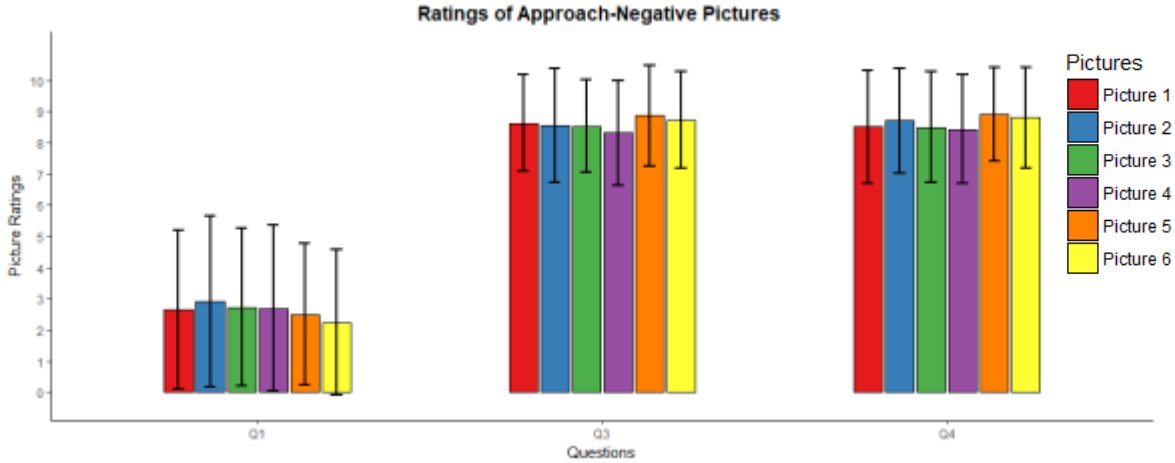
additional information, such as name, age or contact (except the ones who wanted feedback), was asked for.

All students involved in this part of the thesis did not have access to the second part of the current thesis that took place among the students of the University of Lisbon.

6.3.3 Results and Conclusions

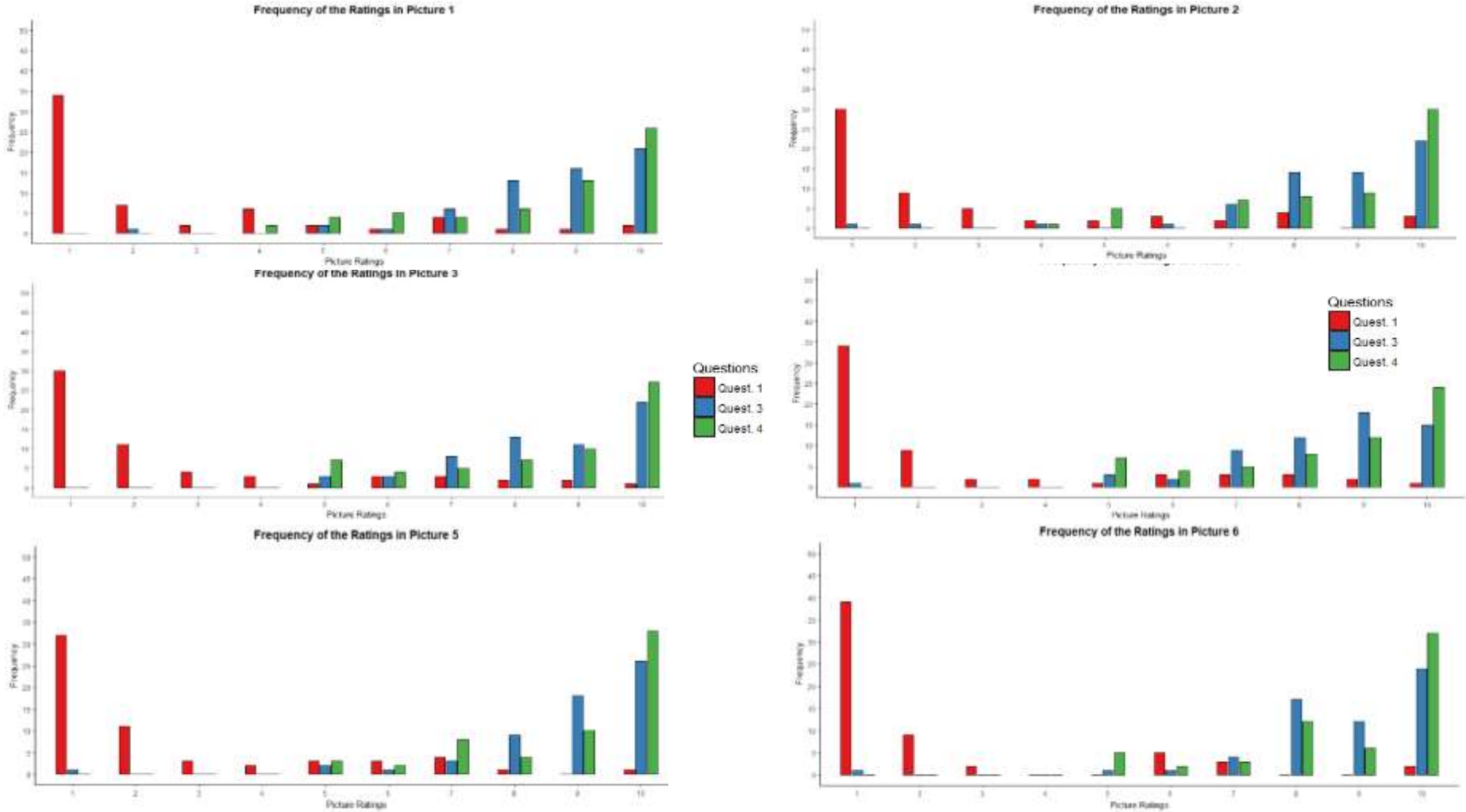


Annexed Figure 13: Ratings of the Neutral-Kitchen Images - Average ratings (y-axis) displayed for each neutral-kitchen image, in the three scale-based questions (x-axis). Q1 - Question Number 1 of the Online Questionnaire - "Quão fácil/difícil é perceber o conteúdo da imagem?" ("How easy/hard it is to understand the image?"); Q3 - "Quão agradável/desagradável é esta imagem para ti, neste momento?" ("How pleasant / unpleasant is this image for you, right now?"); Q4 - "A tua reação a esta imagem é..." ("Your reaction to this image is…?"). The second question, where participants had to describe the image contents in a simple answer, is not shown here.

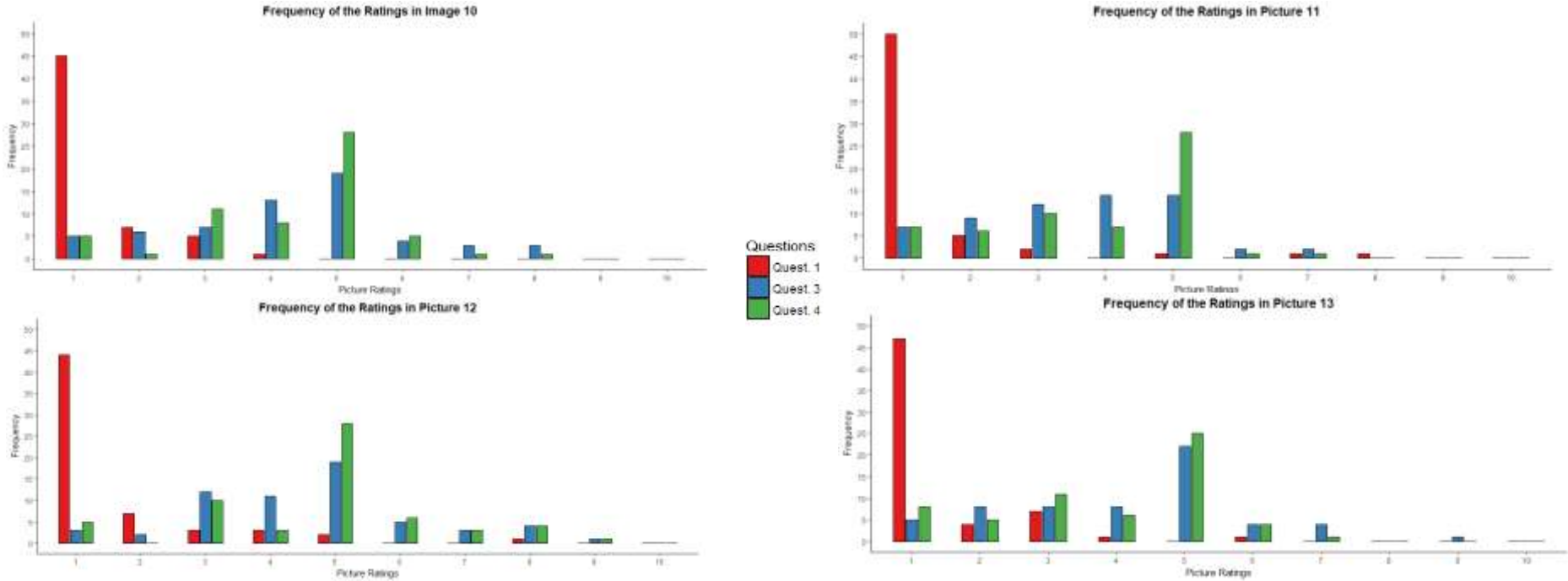


Annexed Figure 14: Ratings of the Negative Images - Average ratings (y-axis) displayed for each neutral-kitchen image, in the three scale-based questions (x-axis). Q1 - Question Number 1 of the Online Questionnaire - "Quão fácil/difícil é perceber o conteúdo da imagem?" ("How easy/hard it is to understand the image?"); Q3 - "Quão agradável/desagradável é esta imagem para ti, neste momento?" ("How pleasant/unpleasant is this image for you, right now?"); Q4 - "A tua reação a esta imagem é..." ("Your reaction to this image is…?"). The second question, where participants had to describe the image contents in a short answer, is not represented in this shown here.

Visual inspection of figures 13 and 14 indicate that all images were considered as easy to understand, since in the *Comprehensibility* question all of them were rated close to the "1" mark (left set of bars, above *Q1*). Regarding the other two questions, visual inspection of the figures also indicated that the negative images were experienced as being more unpleasant and eliciting a strong avoidance response, since in the *Strength of Content* and *Reaction Elicited* questions these images were rated close to the "10" mark (middle and right set of bars, above *Q3* and *Q4*, respectively). As for the neutral-kitchen images, visual inspection indicated that these images were experienced as neither too unpleasant nor too pleasant and did not seem to elicit any specific reaction, since in the *Strength of Content* and *Reaction Elicited* questions these images were rated close to the mark "5". However, visual inspection of the figures indicated that in each category the images were rated very similar, as evidenced by error bars overlap in each question, which made any intent to distinguish them - in order to select the ones whose features were more salient- futile. Thus, to differentiate the images in more detail taking into account that they seemed to be all easy to recognize (Figures 13 and 14, on Q1)- it was decided to analyse the frequency of the ratings specifically on the third and fourth questions (Q3 and Q4, respectively) in each image.



Annexed Figure 15: Frequency of the Ratings for the Negative Images - Frequency (y-axis) of each rating scores (1-10; x-axis) for the questions Q1 (red) - "How easy/hard it is to understand the image?"; Q3 (blue) - "How pleasant/unpleasant is this image for you, right now?"; Q4 (green) - "Your reaction to this image is…(Approach/Avoid)?", for the negative and neutral categories.



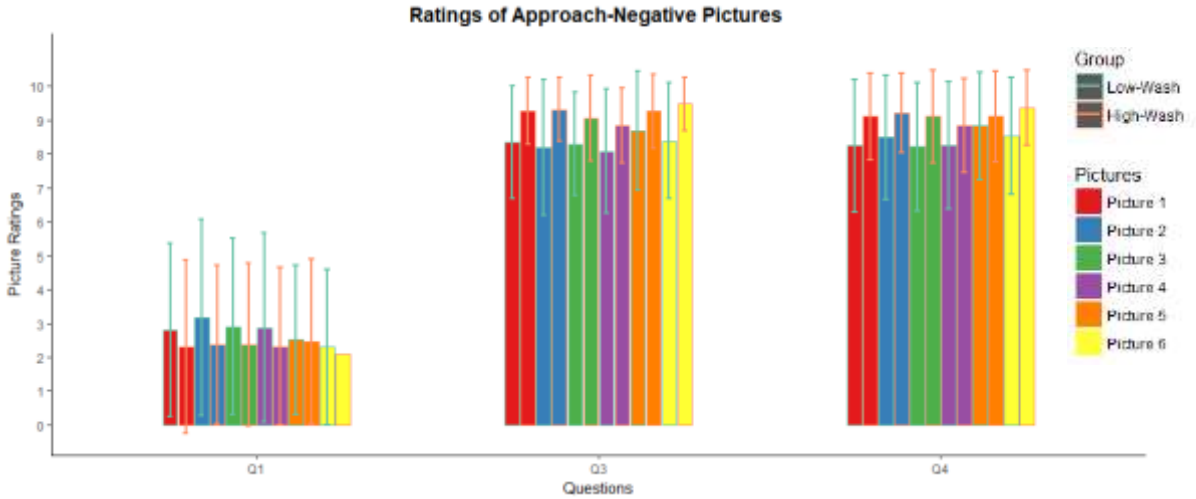
Annexed Figure 16: Frequency of the Ratings for the Neutral Images - Frequency (y-axis) of each rating scores (1-10; x-axis) for the questions Q1 (red) - "How easy/hard it is to understand the image?"; Q3 (blue) - "How pleasant/unpleasant is this image for you, right now?"; Q4 (green) - "Your reaction to this image is…(Approach/Avoid)?", for the negative and neutral categories

With regards to the negative images, visual inspection of figure 15 showed that participants rated the negative images number 2, 5 and 6 as being more unpleasant (see the high number of participants who rated 10 (maximum score) on the third ("*Very Unpleasant"*) and fourth ("*Avoid it as far as possibl*e") question [*Image 2* - Q3, answer "*10*": n = 26/60; Q4, answer "*10*": n = 30/60; *Image 5* - Q3, answer "*10*": n = 26/60; Q4, answer "*10*": n = 33/60; *Image 6* - Q3, answer "*10*": n = 24/60; Q4, answer "*10*": n = 33/60], relative to the negative images number 1, 3 and 4 [*Image 1* - Q3, answer "*10*": n = 21/60; Q4, answer "*10*": n = 26/60; *Image 3* - Q3, answer "*10*": n = 22/60; Q4, answer "*10*": n = 26/60; *Image 4* - Q3, answer "*10*": n = 15/60; Q4, answer "*10*": n = 25/60]).

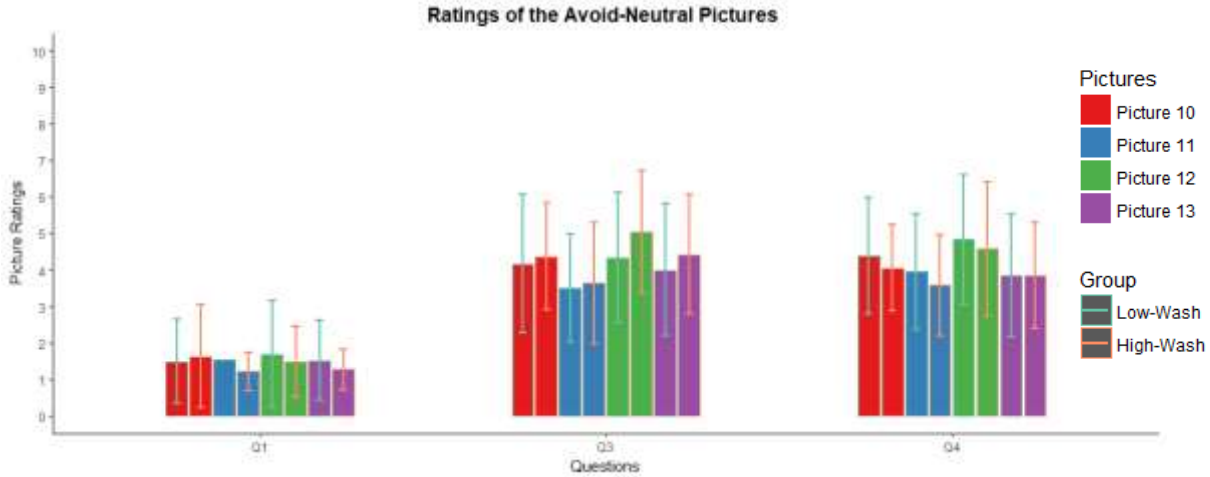
Concerning the neutral-images, the purpose, to briefly remind to reader, was to simply confirm their neutral content. Indeed, visual inspection of figure 16 showed that participants perceived all neutral kitchen pictures in an equally indifferent manner, as seen by the high number of participants that rated "*5*", the equidistant mark relative to the two poles, in both the third and the fourth question (*Image 10* - Q3, answer "5": n= 20/60; Q4, answer "5": n=29/60; *Image 11* - Q3, answer "5": n= 15/60; Q4, answer "5": n=28/60; *Image 12* - Q3, answer "5": n= 20/60; Q4, answer "5": n=27/60; *Image 13* - Q3, answer "5": n= 23/60; Q4, answer "5": n=25/60). For this case, we decided to use these four neutral kitchen pictures.

Prior to the online questionnaire, only 6 out of the 8 negative images had been pre-selected due to their more appropriate features. Upon taking into account the results obtained in the online questionnaire, the negative images number 2 and 5 were assigned only to the AAT Assessment and not to the AAT Training due the following reasons: (1) to allocate an equal number of negative and neutral-kitchen images to the AAT Training, i.e., four negative and four neutral-kitchen images (see Figure 8 in section *2.2.1.3 Training Version* in the Methods); (2) to have 2 strong negative images (images 2 and 5) and 2 weak negative pictures (the latter were not included in the questionnaire due to their evident weak features) for the AAT assessment version, in order to analyse training effects to untrained stimuli with different strength of content levels; (3) to have four negative images allocated to the AAT Training with medium content strength, so that all participants trained with similar content. To summarize, negative images number 1, 3, 4 and 6 (medium content) were allocated to the AAT Training, while negative images number 2, 5 (strong content strength) and other two not included in these ratings (weak content strength) allocated for the AAT Assessment.

To further analyse these results, it was decided to perform the same analysis as in Figures 13 and 14, but now with splitting the ratings of each image into two groups: We compared the ratings performed by participants with a Washing Subscale score higher or equal than 4 (*HG*) *versus* participants with a Washing subscale score lower than 4 (*LG*). For this analysis, it was expected for participants in the former group to rate the negative images as being more unpleasant and eliciting a stronger avoidance response than for participants in the latter group, since, in theory, the (dirty, unwashed) content of the negative images would impact the HG in a more significant way due to their higher contamination-related traits. As for the neutral kitchen images, it was not expected any difference between the groups, since they merely served as a control condition.



Annexed Figure 17: Ratings of Negative Images between Groups - Average ratings (y-axis) displayed by each group for each negative image, in the three scale-based questions (x-axis). Blue outline - low fear of contamination trait group (LG). Pink outline - low fear of contamination trait group (HG). Q1 - Question Number 1 of the Online Questionnaire - "How easy/hard it is to understand the image?" ("How easy/hard it is to understand the image?"); Q3 - "How pleasant/unpleasant is this image for you, right now?" ("How pleasant/unpleasant is this image for you, right now?"); Q4 - "Your reaction to this image is…(Approach/Avoid)?" ("Your reaction to this image is…?"). The second question, where participants had to describe the image contents in a short answer, is not shown here.



Annexed Figure 18:Ratings of Neutral Images between Groups - Average ratings (y-axis) displayed by each group for each neutral-kitchen image, in the three scale-based questions (x-axis). Blue outline - low fear of contamination trait group (LG). Pink outline - low fear of contamination trait group (HG). Q1 - Question

Number 1 of the Online Questionnaire - “How easy/hard it is to understand the image?” (“How easy/hard it is to understand the image?”); Q3 - “How pleasant/unpleasant is this image for you, right now?” (“How pleasant/unpleasant is this image for you, right now?”); Q4 - “Your reaction to this image is…(Approach/Avoid)?” (“Your reaction to this image is…?”). The second question, where participants had to describe the image contents in a short answer, is not represented in this shown here.

Visual inspection of the figure 17 and18 revealed no major group differences in the ratings for the negative and neutral-kitchen images. However, in the negative images (figure 17) one can see a tendency towards of what was expected, specifically in the third and fourth questions: participants of the HG seem to have rated these images as being slightly more unpleasant and eliciting a slightly stronger avoidance reaction, compared to the LG. This was a first support for the intention to select participants for the main study according to their level of fear of contamination. However, statistical analyses were not performed here, since in this pre-study there was a pronounced difference in the number of participants between groups ($N_{High\text{-}Wash}$= 18; $N_{Low\text{-}Wash}$ = 42). With regards to the neutral-kitchen images (Figure 18), visual inspection did not indicate major groups differences in the ratings.

Lastly, an estimation of the frequency of participants with a score above or equal to 4 on the Washing Subscale of the OCI-R was performed to confirm the initial estimation made (see section *6.1 Estimation of the Number of Participants for Screening* in the Annexes). To do so, the number of participants above the Washing Subscale cut-off of 4 (19) was divided by with the total number of participants who filled in the online questionnaire (60). The result was approximately 32%, roughly three times higher than the initial estimation (10%) we obtained with the data of the previous study. This result might be explained due an initial interest in the topic at hand, as it was explicitly stated in the email that was sent to each student.

Taken together, the results of the online questionnaire demonstrated that the new set of negative and neutral kitchen pictures were perceived and rated as having appropriate characteristics that made them a good choice to implement them as part of the AAT stimuli.

6.4 Psychometric Analysis of the OCI-R and BSI questionnaires

In order to briefly compare the quality of the Portuguese OCI-R and BSI versions translated at the laboratory and used in the current thesis *versus* the adapted Portuguese and original English versions, Confirmatory Factor Analyses (CFA) and internal consistency (Cronbach's Alpha) were assessed between the three versions.

To perform the comparisons, the data gathered from all 343 participants that filled in the online questionnaire, which contained the OCI-R and BSI, was used to calculated the different indexes for the current study (see Annexed Tables 1, 2, 3 and 4 below). The missing values in the BSI were handled separately in the following manner for each participant: for every missing value in a participant, the mean of the rest of the items' scores in a subscale was calculated and the value obtained was assigned to the missing value of the respective subscale, in the respective participant. There were no missing values in the OCI-R data.

Both the CFA and the Alphas were calculated using the programming language R version 1.2.1335 in the open source software R Studio, using the packages *psych* and *sem*. It is important to note that particularly for the CFA, the inherent number of subscales and respective items in the OCI-R and BSI scale was used as input, as this analysis was meant to check the "connections" between subscales and respective items.

Annexed Table 1: Confirmatory Analysis Factor in Different OCI-R Versions. GFI-Goodness of Fit Index. SRMR-Standard Root Mean Square Residual. RMSEA-Root Mean Square Error of Approximation. ***Faria and Cardoso (2017); **Foa et al. 2002; *Huppert et al. 2007.

Indexes	Current Study	Portuguese Version***	English Version
Chi-Square	271.68 (120, N = 343), p < .001	289.44 (120, N = 519), p < .001	216.4(121, N = 186), p < .0001* 351.0(138, N = 338), p < .01**
CFI	0.929	0.942	0.95* 0.946**
GFI	0.922	0.925	0.89* 0.897**
SRMR	0.073	0.048	0.099* 0.07**
RMSEA	0.061	0.052	0.06* 0.067**
90% CI Boundries for RMSEA	0.05-0.07	Upper Limit: 0.06	0.05-0.08*

Annexed Table 2: Loading Factors and Internal Consistency (Cronbach's Alpha) for the Washing Subscale and Its Items in Different OCI-R Versions***Faria & Cardoso (2017); **Foa et al. 2002; *Huppert et al. 2007.

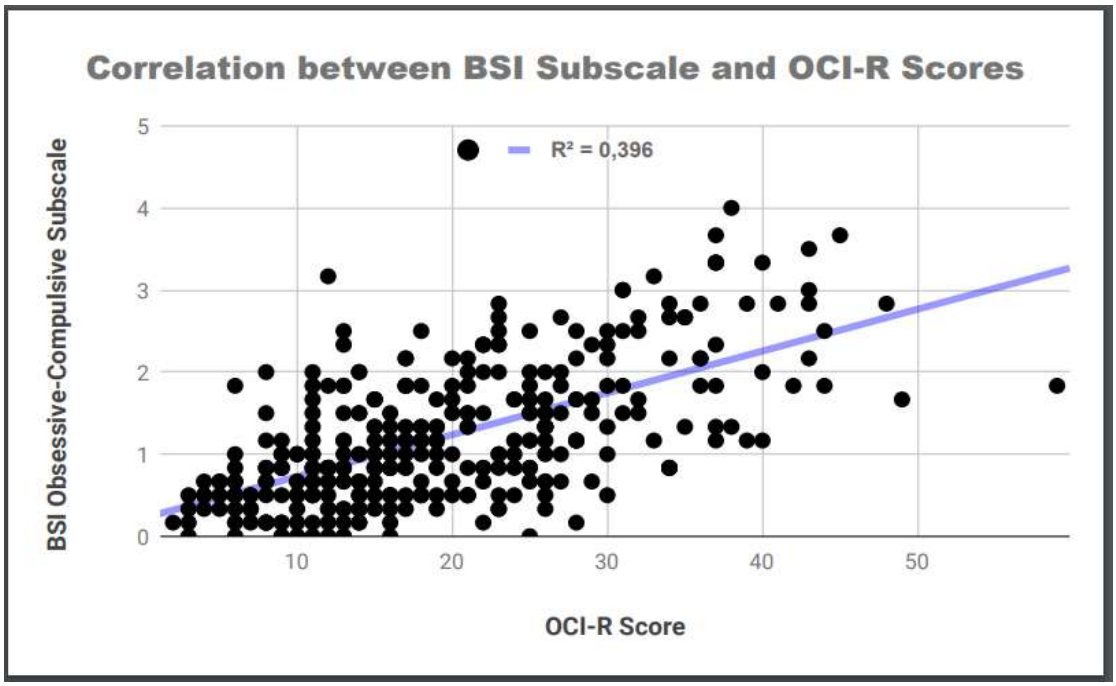
Items	Current Study	Portuguese Version***	English Version
OCI-R Item 5	0.60	0.73	0.77**
OCI-R Item 11	0.83	0.72	0.78**
OCI-R Item 17	0.65	0.68	0.77**
Alpha Value for the Washing Subscale	0.73	0.75	0.69* 0.86**
Alpha Value for the entire OCI-R	0.87	0.891	0.84*

Annexed Table 3: Confirmatory Factor Analysis in Different BSI Versions- CFI-Comparative Fit Index. GFI-Goodness of Fit Index. SRMR-Standard Root Mean Square Residual. RMSEA-Root Mean Square Error of Approximation. *Canavarro (1999); **Derogratis (1983).

Indexes	Current Study	Portuguese Version*	English Version**
Chi-Square	3452.0 (1280, N = 343), p < .001	Not Reported	Not Reported
CFI	0.810	Not Reported	Not Reported
GFI	0.71	Not Reported	Not Reported
SRMR	0.058	Not Reported	Not Reported
RMSEA	0.070	Not Reported	Not Reported

Annexed Table 4: Factor Loadings and Internal Consistency (Cronbach's alpha) for the Obsessive-Compulsive Subscale and Its Items in Different BSI Versions *Canavarro (1999); **Derogratis (1983).

Indexes	Current Study	Portuguese Version*	English Version**
BSI Item 5	0.71	Not reported	0.62
BSI Item 15	0.82	Not reported	0.37
BSI Item 26	0.58	Not reported	0.48
BSI Item 27	0.95	Not reported	0.43
BSI Item 32	0.73	Not reported	0.53
BSI Item 36	0.81	Not reported	0.53
Alpha Value for O-C Subscale	0.86	0.77	0.85
Alpha Value for Total BSI Score	0.97	Not reported	Not reported



Annexed Figure 19: Correlation between the Obsessive-Compulsive BSI Subscale and the Total OCI-R Scores. Each dot corresponds to one subject.